Data Access and Discovery
October 13, 2022

Data Access and Discovery : BHF Data Science Centre

BHF DSC Health Data Science Team

✉ john.nolan@hdruk.ac.uk

British Heart Foundation
Data Science Centre
Led by Health Data Research UK

HDRUK
Health Data Research UK

What is the BHF Data Science Centre?

Partnership

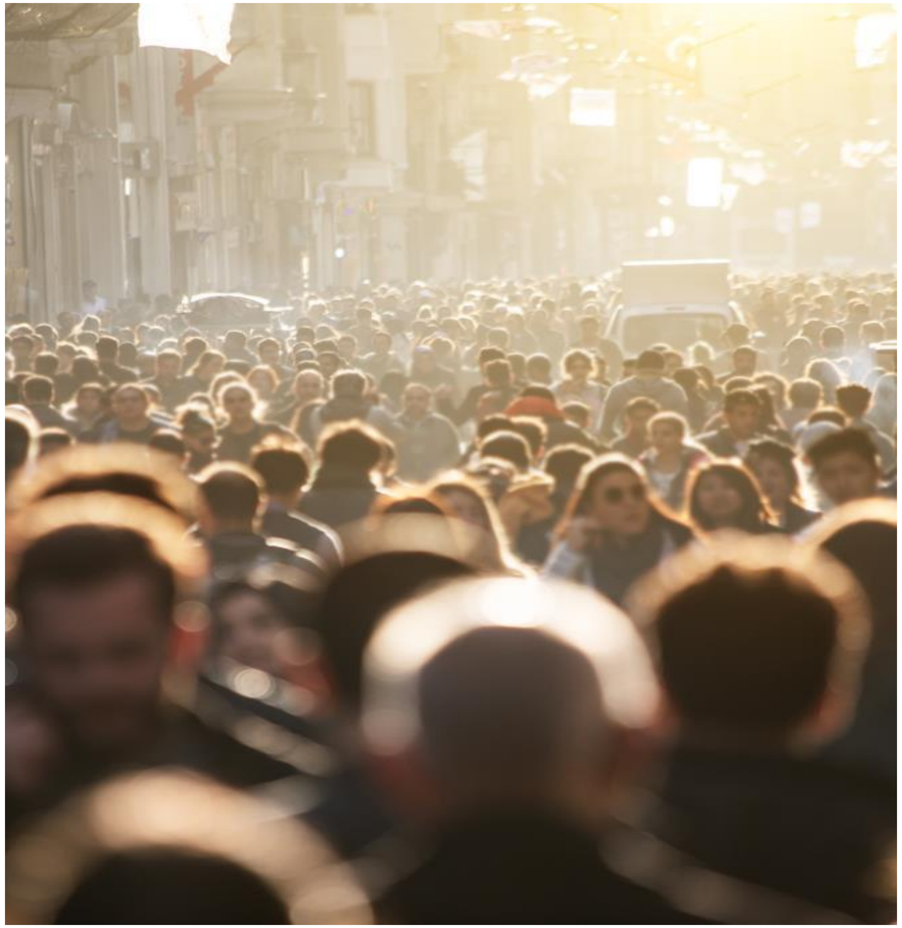
HDRUK
Health Data Research UK

British Heart Foundation

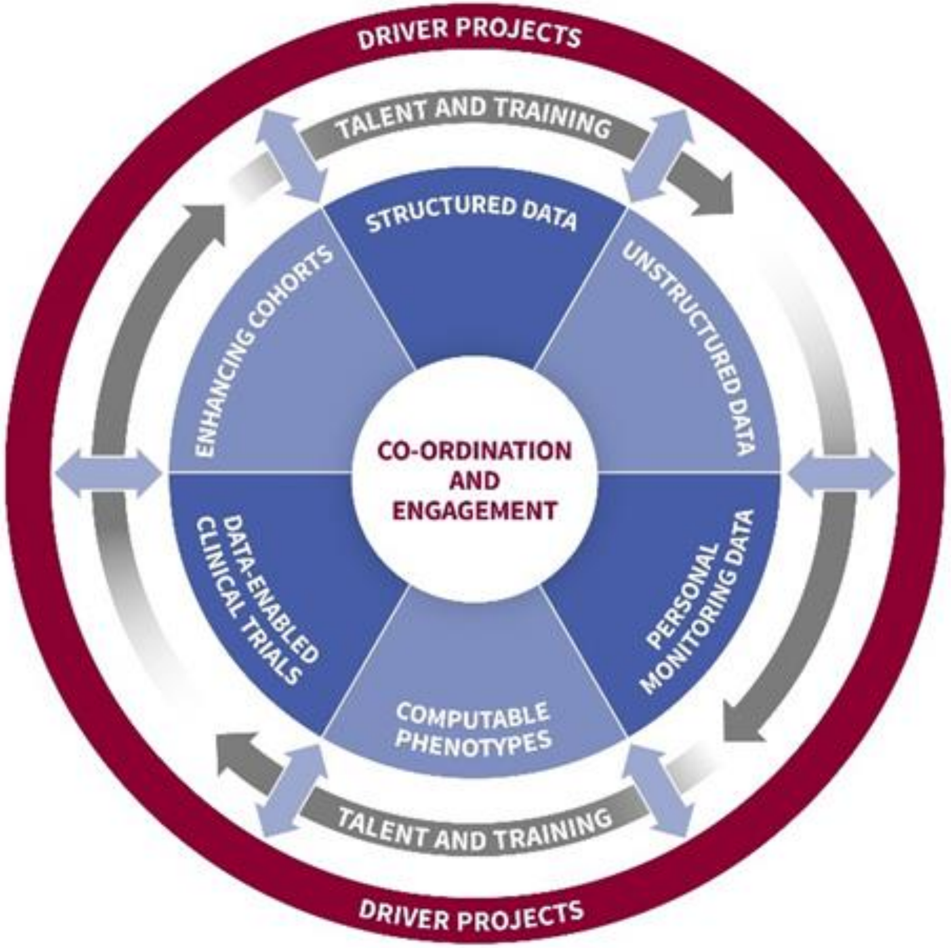
- Officially launched on 1 January 2020
- Partnership between HDR UK and BHF
- Funding from BHF: £10M over 5 years

Our vision

To improve the cardiovascular health of the nation using the power of large-scale data and advanced analytics across the UK



Themes and cross-cutting activities



6 thematic areas:

- Better access to and use of nationally-collated, structured, coded data
- Better access to and use of unstructured health data
- Personal monitoring data
- Computable cardiovascular phenotypes
- Enhancing cohorts
- Data-enabled clinical trials

Diabetes Data Science Catalyst

DiABETES UK
KNOW DIABETES. FIGHT DIABETES.

3 cross-cutting activities:

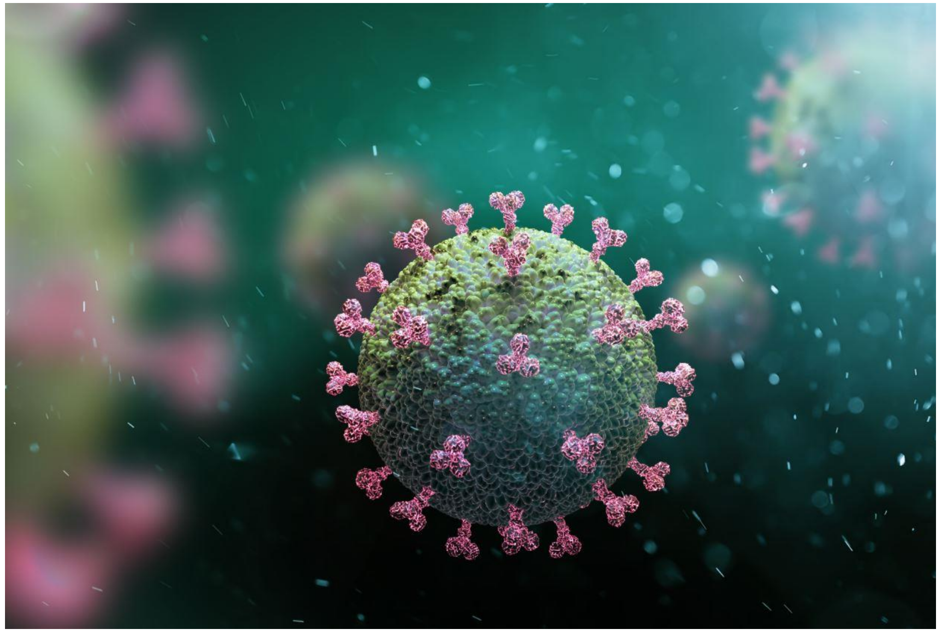
- Co-ordination and Engagement
- Talent and Training
- Driver projects

CVD-COVID-UK/COVID-IMPACT: aims

British Heart Foundation Data Science Centre
Led by Health Data Research UK

CVD-COVID-UK

- Aims to understand the relationship between COVID-19 and cardiovascular diseases such as heart attack, heart failure, stroke, and blood clots in the lungs
- Achieved through analyses of de-identified, pseudonymised, linked, nationally collated healthcare data sources in trusted research environments (TREs) across the four nations of the UK



COVID-IMPACT

- Builds on the success of CVD-COVID-UK by broadening the scope of the programme to **all** COVID-related research (currently using data in NHS Digital's TRE for England only)
- Helps to support research projects from the wider community, including for the Data & Connectivity National Core Study

CVD-COVID-UK/COVID-IMPACT Consortium in numbers

>280 members

>50 institutions

>90 analysts

34 projects

3 national TREs

67 datasets

3 publications

7 preprints

>60 studies in progress

https://www.hdruk.ac.uk/projects/cvd-covid-uk-project/

CVD-COVID-UK/COVID-IMPACT TRE Dataset Provisioning Dashboard: 28/09/22

British Heart Foundation Data Science Centre
Led by Health Data Research UK

Links: Innovation Gateway TRE Dataset/Access Request   Innovation Gateway Collection   GitHub   Paper on the power of data linkage

Nation / Population size	ENGLAND / 57 million	SCOTLAND / 5.5 million	WALES / 3.2 million
TRE	NHS Digital's TRE service for England	National Data Safe Haven	SAIL Databank
Users / Institutions	76 users / 10 institutions	16 users / 6 institutions	33 users / 12 institutions
Datasets	33 requested / 26 provisioned	18 requested / 16 provisioned	34 requested / 30 provisioned
Comments	• NICOR NACSA/NACRM provisioned • Maternity Services in the pipeline	• SMR02 to be requested	• ONS COVID-19 Infection Survey available, subject to additional approvals
Primary Care	🟩 • GDPPR	🟩 • Primary Care	🟩 • General Practice Monthly/Daily COVID
Secondary Care	🟩 • HES (Admitted Patient Care, Outpatient, Critical Care, Accident & Emergency) 🟩 • SUS 🟩 • Uncurated Low Latency Hospital Data 🟩 • Emergency Care Data Set	🟩 • Outpatient Appointments / Attendances - Scottish Morbidity Record (SMR00) 🟩 • General Acute Inpatient and Day Case - Scottish Morbidity Record (SMR01) 🟩 • Accident & Emergency	🟩 • Critical Care Dataset 🟩 • Emergency Department Daily/Monthly 🟩 • Outpatient Dataset for Wales 🟩 • Outpatient Referral Dataset 🟩 • Patient Episode Dataset
Covid-19 Lab Tests	🟩 • SGSS (Pillar 1, 2 – positive results only) 🟩 • Pillar 2 Antigen (positive and negative) 🟩 • Pillar 3 Antibody (positive and negative) 🟧 • Variant strain data (COG-UK)	🟩 • COVID Tests (lab/lighthouse testing) 🟩 • (ECOSS) 🟩 • Variant strain data (COG-UK)	🟩 • LIMS (Pillar 1, 2, 3) 🟧 • ONS COVID-19 Infection Survey* 🟩 • Test, Trace & Protect 🟩 • Shielded People 🟩 • Variant strain data (COG-UK)*
Covid-19 Vaccinations	🟩 • Covid-19 vaccination events 🟩 • Covid-19 vaccination adverse reactions	🟩 • Vaccination Data	🟩 • Covid Vaccination Dataset
Deaths	🟩 • Civil Registry Deaths	🟩 • Deaths	🟩 • Annual District Death Daily/Monthly 🟩 • Consolidated Death Data Source
ITU	🟩 • ICNARC COVID	🟩 • SICSAG Daily, Episodes	🟩 • ICNARC Quarterly/Weekly COVID
ITU/HDU Admissions	🟩 • (COVID-19 SARI-Watch - formerly CHESS)	⬜ • N/A	⬜ • N/A
Prescribing/Dispensing	🟩 • NHS BSA Dispensed Medicines 🟩 • Secondary care prescribed medicines	🟩 • PIS: Dispensed, Prescribed, Paid 🟧 • ePrescribing	🟩 • Wales Dispensing Dataset
NICOR CVD Audits	🟩 • PCI, MINAP, NHFA, NCHDA, NACRM, NACSA 🟧 • TAVI	⬜ • N/A	🟧 • NICOR Audits and Registers (pending approvals)
Stroke Audit	🟩 • SSNAP	🟩 • Scottish Stroke Care Audit (SSCA)	🟧 • HQIP Stroke Audit (pending approvals)
National Vascular Registry	🟧 • NVR	⬜ • NVR (not currently requested)	🟧 • NVR (pending approvals)
Other	🟩 • Improving Access to Psychological Therapies (IAPT v2.0) 🟧 • Maternity Services Data Set 🟧 • Mental Health Data Set 🟧 • Mental Health of Children and Young People 🟧 • Patient Reported Outcome Measures	🟩 • Diabetes Covariates 🟩 • Scottish Renal Registry 🟧 • Maternity Inpatient and Day Case - Scottish Morbidity Record (SMR02)	🟩 • Annual District Birth Extract 🟩 • Care Homes Index 🟩 • Maternity Indicators Dataset 🟩 • Congenital Anomaly Register (CARIS) 🟩 • National Community Child Health 🟩 • ONS Census (2011)* 🟩 • Referral to Treatment Times 🟩 • SAIL Dementia e-Cohort 🟩 • Welsh Ambulance Service Dataset 🟩 • Wales Results Reporting Service 🟩 • Welsh Demographic Service

NORTHERN IRELAND

Access to corresponding datasets to follow

KEY

🟩 Dataset available and actively being used for research purposes
🟧 Dataset requested, but not yet available / pending approvals
⬜ Dataset not requested
* Additional approvals required

DATASET ACRONYMS

- **CHESS:** COVID-19 Hospitalisation in England Surveillance System
- **ECOSS:** Electronic Communication of Surveillance in Scotland
- **GDPPR:** General Practice Extraction Service (GPES) Data for Pandemic Planning and Research
- **HES:** Hospital Episode Statistics
- **HQIP:** Healthcare Quality Improvement Partnership
- **ICNARC:** Intensive Care National Audit and Research Centre
- **LIMS:** Laboratory Information Management System
- **MINAP:** Myocardial Ischaemia National Audit Project
- **NACRM:** National Audit of Cardiac Rhythm Management
- **NACSA:** National Adult Cardiac Surgery Audit
- **NCHDA:** National Congenital Heart Disease Audit
- **NHFA:** National Heart Failure Audit
- **NICOR:** National Institute for Cardiovascular Outcomes Research
- **NIMS:** National Immunisation Management System
- **NVR:** National Vascular Registry
- **PCI:** Percutaneous Coronary Interventions
- **SGSS**: Second Generation Surveillance System
- **SICSAG:** Scottish Intensive Care Society Audit Group
- **SSNAP:** Sentinel Stroke National Audit Programme
- **SUS:** Secondary Uses Service
- **TAVI:** Transcatheter Aortic Valve Implantation

CVD-COVID-UK/COVID-IMPACT Projects

Methods

- Data management and analysis methods
- High-throughput phenotyping approaches
- Improving methods to minimise bias in ethnicity data

Medicines

- Effects of ACE inhibitors & ARBs on COVID-19
- Impact of COVID-19 on managing BP and lipids
- Assessing COVID-19 impact through medicines
- Antipsychotic prescribing during the pandemic and cardiovascular risk in patients with dementia
- Evaluation of antithrombotic use on COVID-19 outcomes
- Repurposing medicines to prevent COVID-19

Others

- COVID-19 infection, vaccination and vascular risk
- Direct and indirect effects of COVID-19 in people with cardiovascular disease
- COVID and cardiovascular disease risk prediction
- Impact of COVID-19 on Congenital Heart Disease (CHD) patients undergoing cardiac surgery
- Influence of multi-morbidity on outcomes of COVID-19
- Impact of COVID infection and vaccination on pregnancy

- Predicting severe COVID-19 in people with rare diseases
  Genomics of multi-morbidity and susceptibility to COVID-19
- Longer-term effects of COVID-19 in non-hospitalised people
- Evaluating how palliative and end of life care teams have responded to COVID-19
- Coronary revascularisation and outcomes before and after the COVID-19 pandemic
- Children admitted to hospital with COVID-19 – risk factors, risk groups and NHS care utilisation
- Understanding the increased risk of severe COVID-19 in people with intellectual & developmental disabilities
- Risks of cardiovascular disease in people with COVID-19 and pre-existing respiratory disease
- Impact of COVID-19 on eye disease
- Impact of COVID-19 on heart failure
- Impact of COVID-19 on people with diabetes

Health Data Science Team

Senior Health Data Scientists
Dr Tom Bolton
John Nolan

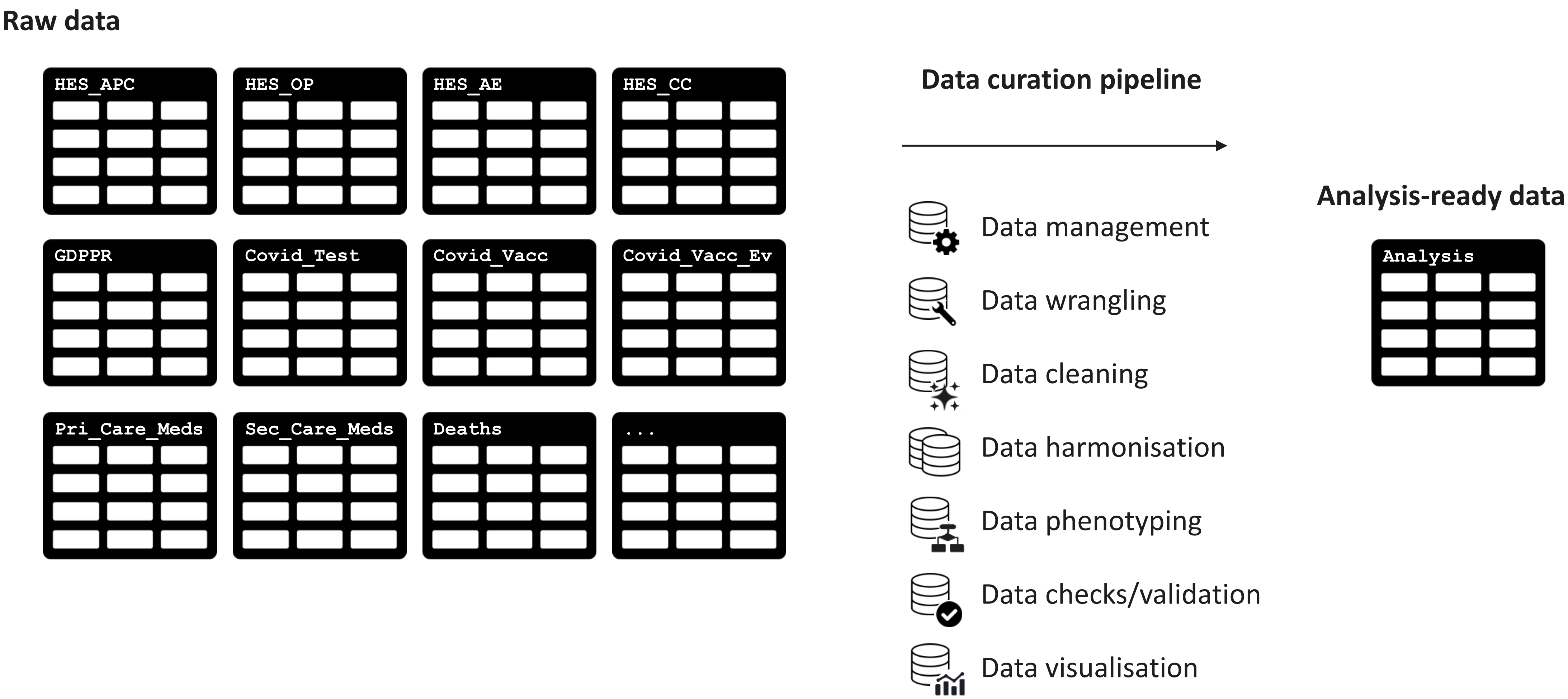
Health Data Scientists
Dr Mehrdad Mizani
Dr Zach Welshman

Early Career Health Data Scientist
Dr Jamie Farrell

September 2021 — May 2022 — June — July — August — September — October

1 x Health Data Scientist +
2 x Early Career Health Data Scientist

What is a data curation pipeline?



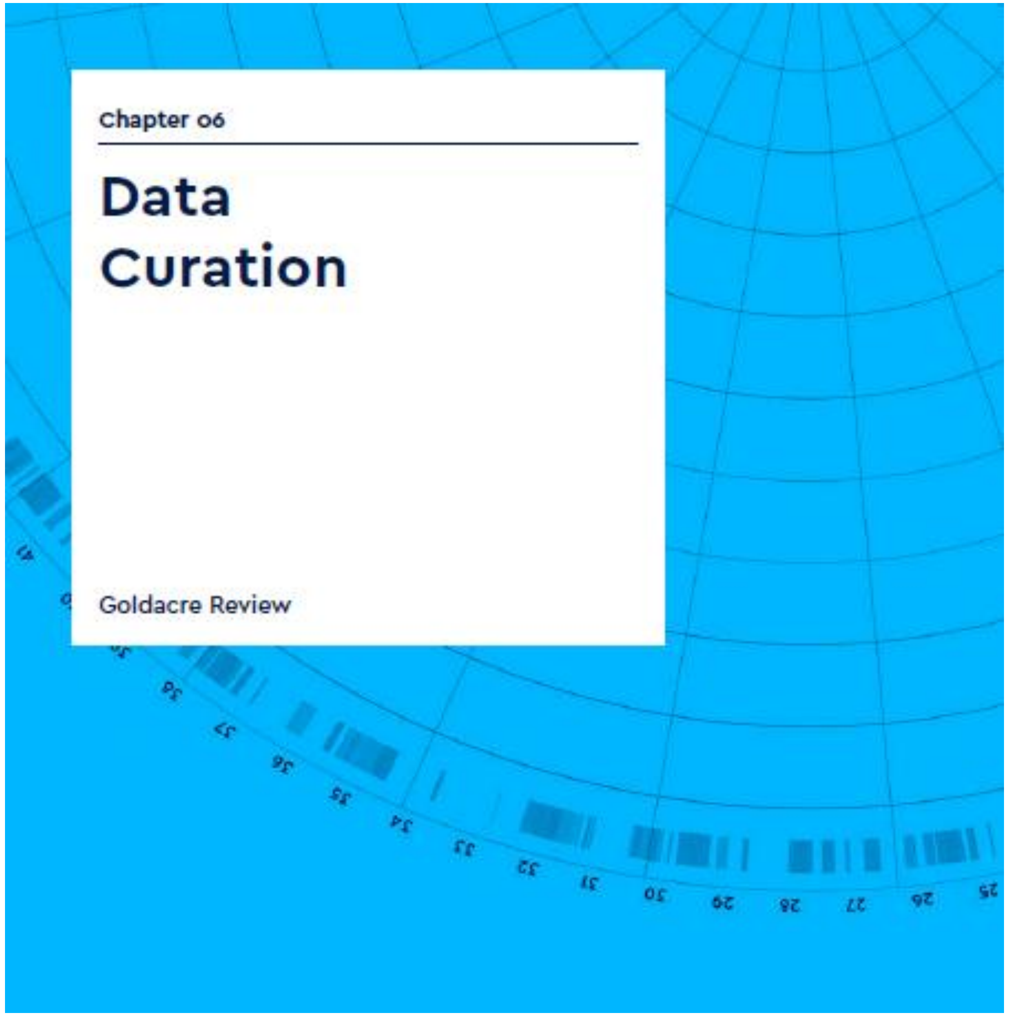
Raw data

Data curation pipeline

- Data management
- Data wrangling
- Data cleaning
- Data harmonisation
- Data phenotyping
- Data checks/validation
- Data visualisation

Analysis-ready data

Motivation



Chapter 06

Data Curation

Goldacre Review

> “It has been estimated that 80% of the work for data science with NHS records is spent on data preparation.”

Resources

Data

> Data notes

> Data dictionary

> Data summary notebooks

> Data insight notebooks

Code

> Demos

> Curated data

> Data curation pipeline functions

Resources

Data

> Data notes

> Data dictionary

> Data summary notebooks

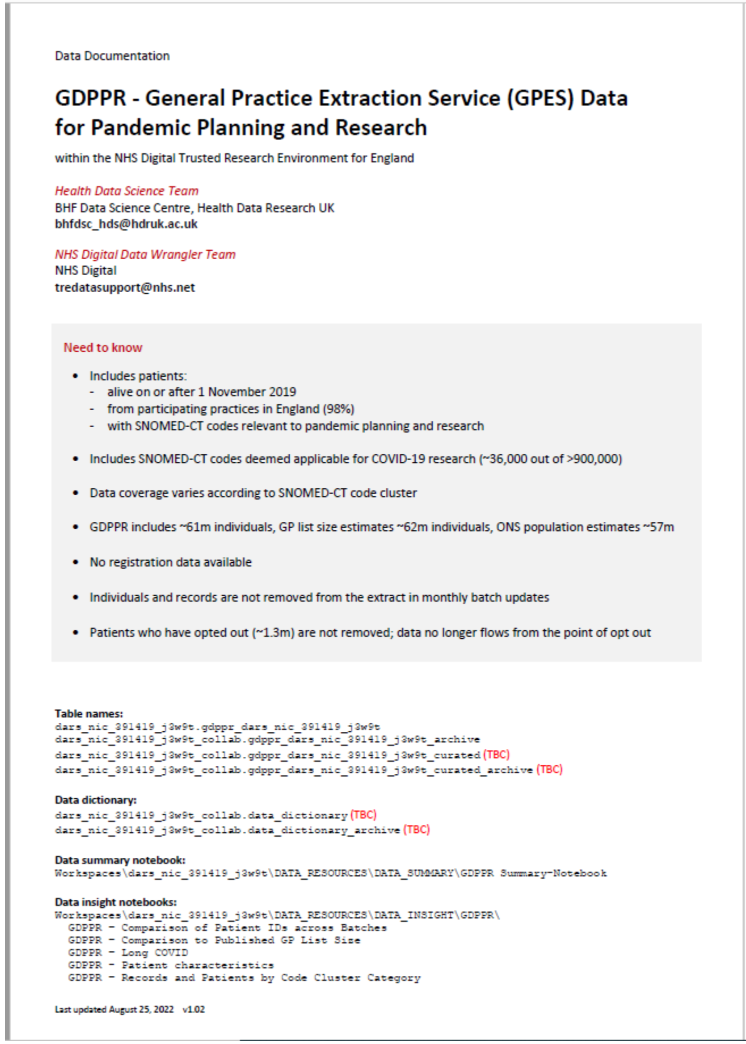
> Data insight notebooks

Code

> Demos

> Curated data

> Data curation pipeline functions



Data Documentation

GDPPR - General Practice Extraction Service (GPES) Data for Pandemic Planning and Research

within the NHS Digital Trusted Research Environment for England

Health Data Science Team
BHF Data Science Centre, Health Data Research UK
bhfdsc_hds@hdruk.ac.uk

NHS Digital Data Wrangler Team
NHS Digital
tredatasupport@nhs.net

Need to know

- Includes patients:
  - alive on or after 1 November 2019
  - from participating practices in England (98%)
  - with SNOMED-CT codes relevant to pandemic planning and research
- Includes SNOMED-CT codes deemed applicable for COVID-19 research (~36,000 out of >900,000)
- Data coverage varies according to SNOMED-CT code cluster
- GDPPR includes ~61m individuals, GP list size estimates ~62m individuals, ONS population estimates ~57m
- No registration data available
- Individuals and records are not removed from the extract in monthly batch updates
- Patients who have opted out (~1.3m) are not removed; data no longer flows from the point of opt out

Table names:

```
dars_nic_391419_j3w9t.gdppr_dars_nic_391419_j3w9t
dars_nic_391419_j3w9t_collab.gdppr_dars_nic_391419_j3w9t_archive
dars_nic_391419_j3w9t_collab.gdppr_dars_nic_391419_j3w9t_curated (TBC)
dars_nic_391419_j3w9t_collab.gdppr_dars_nic_391419_j3w9t_curated_archive (TBC)
```

Data dictionary:

```
dars_nic_391419_j3w9t_collab.data_dictionary (TBC)
dars_nic_391419_j3w9t_collab.data_dictionary_archive (TBC)
```

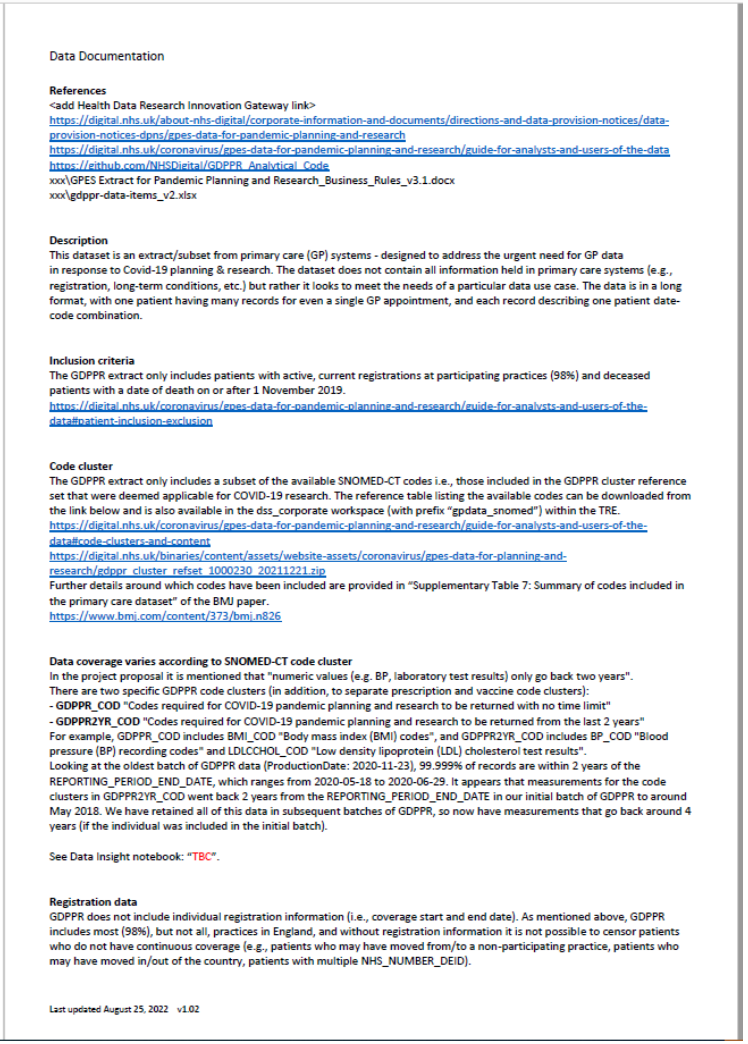
Data summary notebook:

```
Workspaces\dars_nic_391419_j3w9t\DATA_RESOURCES\DATA_SUMMARY\GDPPR Summary-Notebook
```

Data insight notebooks:

```
Workspaces\dars_nic_391419_j3w9t\DATA_RESOURCES\DATA_INSIGHT\GDPPR\
  GDPPR - Comparison of Patient IDs across Batches
  GDPPR - Comparison to Published GP List Size
  GDPPR - Long COVID
  GDPPR - Patient characteristics
  GDPPR - Records and Patients by Code Cluster Category
```

Last updated August 25, 2022 v1.02



Data Documentation

References
<add Health Data Research Innovation Gateway link>
https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/directions-and-data-provision-notices/data-provision-notices-dpns/gpes-data-for-pandemic-planning-and-research
https://digital.nhs.uk/coronavirus/gpes-data-for-pandemic-planning-and-research/guide-for-analysts-and-users-of-the-data
https://github.com/NHSDigital/GDPPR_Analytical_Code
xxx\GPES Extract for Pandemic Planning and Research_Business_Rules_v3.1.docx
xxx\gdppr-data-items_v2.xlsx

Description
This dataset is an extract/subset from primary care (GP) systems - designed to address the urgent need for GP data in response to Covid-19 planning & research. The dataset does not contain all information held in primary care systems (e.g., registration, long-term conditions, etc.) but rather it looks to meet the needs of a particular data use case. The data is in a long format, with one patient having many records for even a single GP appointment, and each record describing one patient date-code combination.

Inclusion criteria
The GDPPR extract only includes patients with active, current registrations at participating practices (98%) and deceased patients with a date of death on or after 1 November 2019.
https://digital.nhs.uk/coronavirus/gpes-data-for-pandemic-planning-and-research/guide-for-analysts-and-users-of-the-data#patient-inclusion-exclusion

Code cluster
The GDPPR extract only includes a subset of the available SNOMED-CT codes i.e., those included in the GDPPR cluster reference set that were deemed applicable for COVID-19 research. The reference table listing the available codes can be downloaded from the link below and is also available in the dss_corporate workspace (with prefix "gpdata_snomed") within the TRE.
https://digital.nhs.uk/coronavirus/gpes-data-for-pandemic-planning-and-research/guide-for-analysts-and-users-of-the-data#code-clusters-and-content
https://digital.nhs.uk/binaries/content/assets/website-assets/coronavirus/gpes-data-for-planning-and-research/gdppr_cluster_refset_1000230_20211221.zip
Further details around which codes have been included are provided in "Supplementary Table 7: Summary of codes included in the primary care dataset" of the BMJ paper.
https://www.bmj.com/content/373/bmj.n826

Data coverage varies according to SNOMED-CT code cluster
In the project proposal it is mentioned that "numeric values (e.g. BP, laboratory test results) only go back two years".
There are two specific GDPPR code clusters (in addition, to separate prescription and vaccine code clusters):
- **GDPPR_COD** "Codes required for COVID-19 pandemic planning and research to be returned with no time limit"
- **GDPPR2YR_COD** "Codes required for COVID-19 pandemic planning and research to be returned from the last 2 years"

For example, GDPPR_COD includes BMI_COD "Body mass index (BMI) codes", and GDPPR2YR_COD includes BP_COD "Blood pressure (BP) recording codes" and LDLCCHOL_COD "Low density lipoprotein (LDL) cholesterol test results".
Looking at the oldest batch of GDPPR data (ProductionDate: 2020-11-23), 99.999% of records are within 2 years of the REPORTING_PERIOD_END_DATE, which ranges from 2020-05-18 to 2020-06-29. It appears that measurements for the code clusters in GDPPR2YR_COD went back 2 years from the REPORTING_PERIOD_END_DATE in our initial batch of GDPPR to around May 2018. We have retained all of this data in subsequent batches of GDPPR, so now have measurements that go back around 4 years (if the individual was included in the initial batch).

See Data Insight notebook: "TBC".

Registration data
GDPPR does not include individual registration information (i.e., coverage start and end date). As mentioned above, GDPPR includes most (98%), but not all, practices in England, and without registration information it is not possible to censor patients who do not have continuous coverage (e.g., patients who may have moved from/to a non-participating practice, patients who may have moved in/out of the country, patients with multiple NHS_NUMBER_DEID).

Last updated August 25, 2022 v1.02

Resources

Data

> Data notes

> Data dictionary

> Data summary notebooks

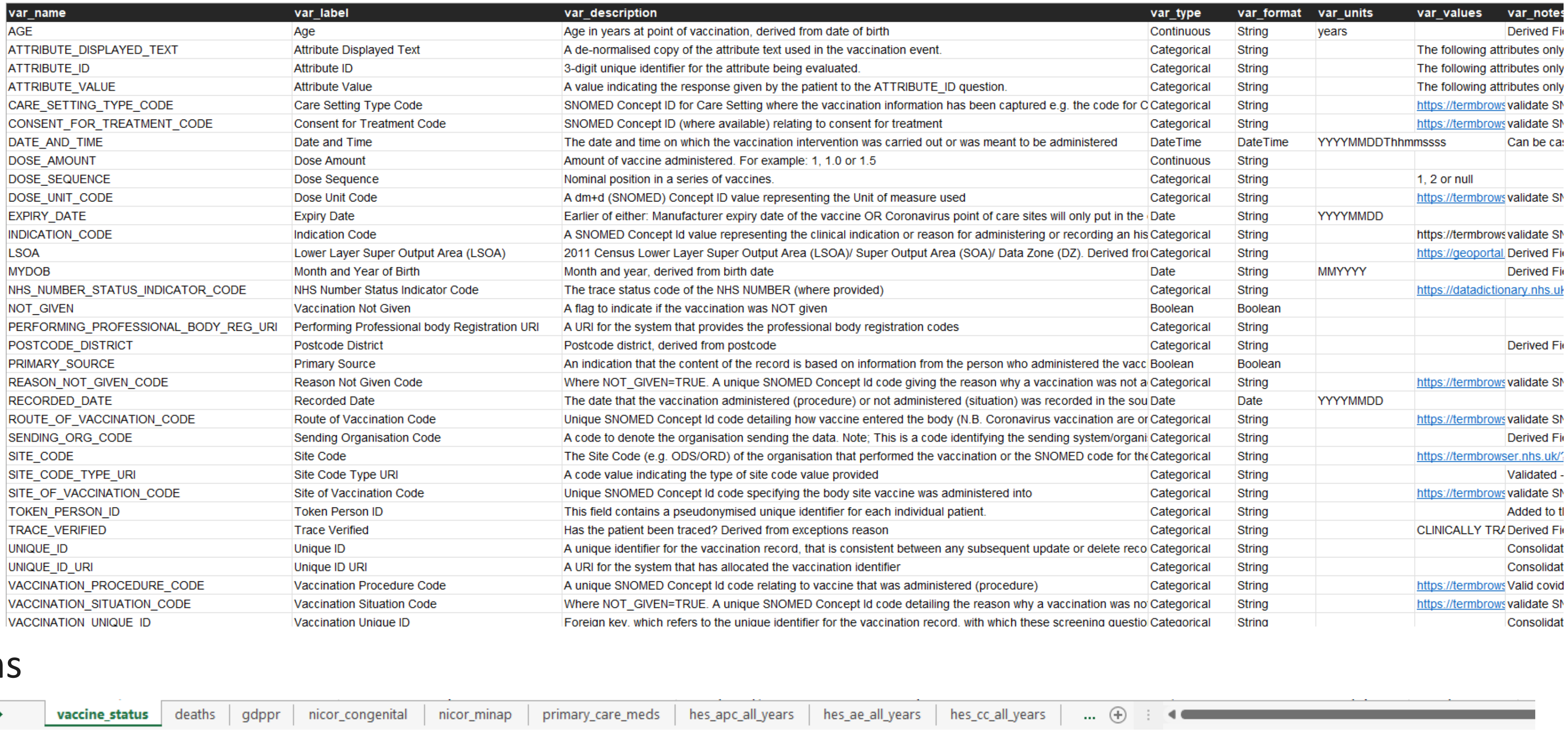
> Data insight notebooks

Code

> Demos

> Curated data

> Data curation pipeline functions



var_name	var_label	var_description	var_type	var_format	var_units	var_values	var_notes
AGE	Age	Age in years at point of vaccination, derived from date of birth	Continuous	String	years		Derived Fi
ATTRIBUTE_DISPLAYED_TEXT	Attribute Displayed Text	A de-normalised copy of the attribute text used in the vaccination event.	Categorical	String		The following attributes only	
ATTRIBUTE_ID	Attribute ID	3-digit unique identifier for the attribute being evaluated.	Categorical	String		The following attributes only	
ATTRIBUTE_VALUE	Attribute Value	A value indicating the response given by the patient to the ATTRIBUTE_ID question.	Categorical	String		The following attributes only	
CARE_SETTING_TYPE_CODE	Care Setting Type Code	SNOMED Concept ID for Care Setting where the vaccination information has been captured e.g. the code for C	Categorical	String		https://termbrows	validate SN
CONSENT_FOR_TREATMENT_CODE	Consent for Treatment Code	SNOMED Concept ID (where available) relating to consent for treatment	Categorical	String		https://termbrows	validate SN
DATE_AND_TIME	Date and Time	The date and time on which the vaccination intervention was carried out or was meant to be administered	DateTime	DateTime	YYYYMMDDThhmmssss		Can be ca
DOSE_AMOUNT	Dose Amount	Amount of vaccine administered. For example: 1, 1.0 or 1.5	Continuous	String			
DOSE_SEQUENCE	Dose Sequence	Nominal position in a series of vaccines.	Categorical	String		1, 2 or null	
DOSE_UNIT_CODE	Dose Unit Code	A dm+d (SNOMED) Concept ID value representing the Unit of measure used	Categorical	String		https://termbrows	validate SN
EXPIRY_DATE	Expiry Date	Earlier of either: Manufacturer expiry date of the vaccine OR Coronavirus point of care sites will only put in the	Date	String	YYYYMMDD		
INDICATION_CODE	Indication Code	A SNOMED Concept Id value representing the clinical indication or reason for administering or recording an his	Categorical	String		https://termbrows	validate SN
LSOA	Lower Layer Super Output Area (LSOA)	2011 Census Lower Layer Super Output Area (LSOA)/ Super Output Area (SOA)/ Data Zone (DZ). Derived fro	Categorical	String		https://geoportal.	Derived Fi
MYDOB	Month and Year of Birth	Month and year, derived from birth date	Date	String	MMYYYY		Derived Fi
NHS_NUMBER_STATUS_INDICATOR_CODE	NHS Number Status Indicator Code	The trace status code of the NHS NUMBER (where provided)	Categorical	String		https://datadictionary.nhs.uk	
NOT_GIVEN	Vaccination Not Given	A flag to indicate if the vaccination was NOT given	Boolean	Boolean			
PERFORMING_PROFESSIONAL_BODY_REG_URI	Performing Professional body Registration URI	A URI for the system that provides the professional body registration codes	Categorical	String			
POSTCODE_DISTRICT	Postcode District	Postcode district, derived from postcode	Categorical	String			Derived Fi
PRIMARY_SOURCE	Primary Source	An indication that the content of the record is based on information from the person who administered the vacc	Boolean	Boolean			
REASON_NOT_GIVEN_CODE	Reason Not Given Code	Where NOT_GIVEN=TRUE. A unique SNOMED Concept Id code giving the reason why a vaccination was not a	Categorical	String		https://termbrows	validate SN
RECORDED_DATE	Recorded Date	The date that the vaccination administered (procedure) or not administered (situation) was recorded in the sou	Date	Date	YYYYMMDD		
ROUTE_OF_VACCINATION_CODE	Route of Vaccination Code	Unique SNOMED Concept Id code detailing how vaccine entered the body (N.B. Coronavirus vaccination are or	Categorical	String		https://termbrows	validate SN
SENDING_ORG_CODE	Sending Organisation Code	A code to denote the organisation sending the data. Note; This is a code identifying the sending system/organi	Categorical	String			Derived Fi
SITE_CODE	Site Code	The Site Code (e.g. ODS/ORD) of the organisation that performed the vaccination or the SNOMED code for the	Categorical	String		https://termbrowser.nhs.uk/	
SITE_CODE_TYPE_URI	Site Code Type URI	A code value indicating the type of site code value provided	Categorical	String			Validated -
SITE_OF_VACCINATION_CODE	Site of Vaccination Code	Unique SNOMED Concept Id code specifying the body site vaccine was administered into	Categorical	String		https://termbrows	validate SN
TOKEN_PERSON_ID	Token Person ID	This field contains a pseudonymised unique identifier for each individual patient.	Categorical	String			Added to t
TRACE_VERIFIED	Trace Verified	Has the patient been traced? Derived from exceptions reason	Categorical	String		CLINICALLY TRA	Derived Fi
UNIQUE_ID	Unique ID	A unique identifier for the vaccination record, that is consistent between any subsequent update or delete reco	Categorical	String			Consolidat
UNIQUE_ID_URI	Unique ID URI	A URI for the system that has allocated the vaccination identifier	Categorical	String			Consolidat
VACCINATION_PROCEDURE_CODE	Vaccination Procedure Code	A unique SNOMED Concept Id code relating to vaccine that was administered (procedure)	Categorical	String		https://termbrows	Valid covid
VACCINATION_SITUATION_CODE	Vaccination Situation Code	Where NOT_GIVEN=TRUE. A unique SNOMED Concept Id code detailing the reason why a vaccination was no	Categorical	String		https://termbrows	validate SN
VACCINATION_UNIQUE_ID	Vaccination Unique ID	Foreign key, which refers to the unique identifier for the vaccination record, with which these screening questio	Categorical	String			Consolidat

Resources

Data

> Data documentation

> Data dictionary

> Data summary notebooks

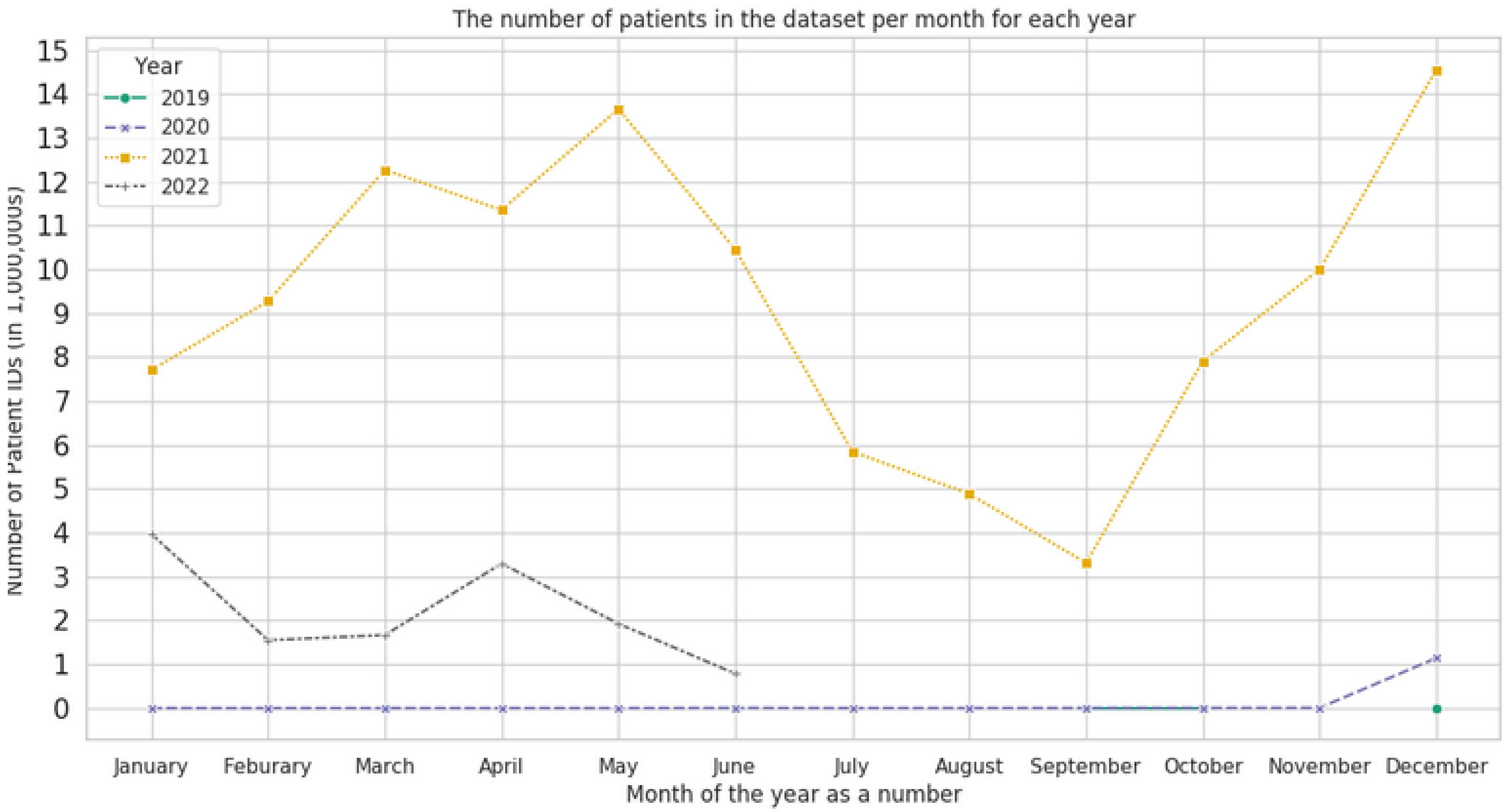
> Data insight notebooks

Code

> Demos

> Curated data

> Data curation pipeline functions



Resources

Data

> Data documentation

> Data dictionary

> **Data summary notebooks**

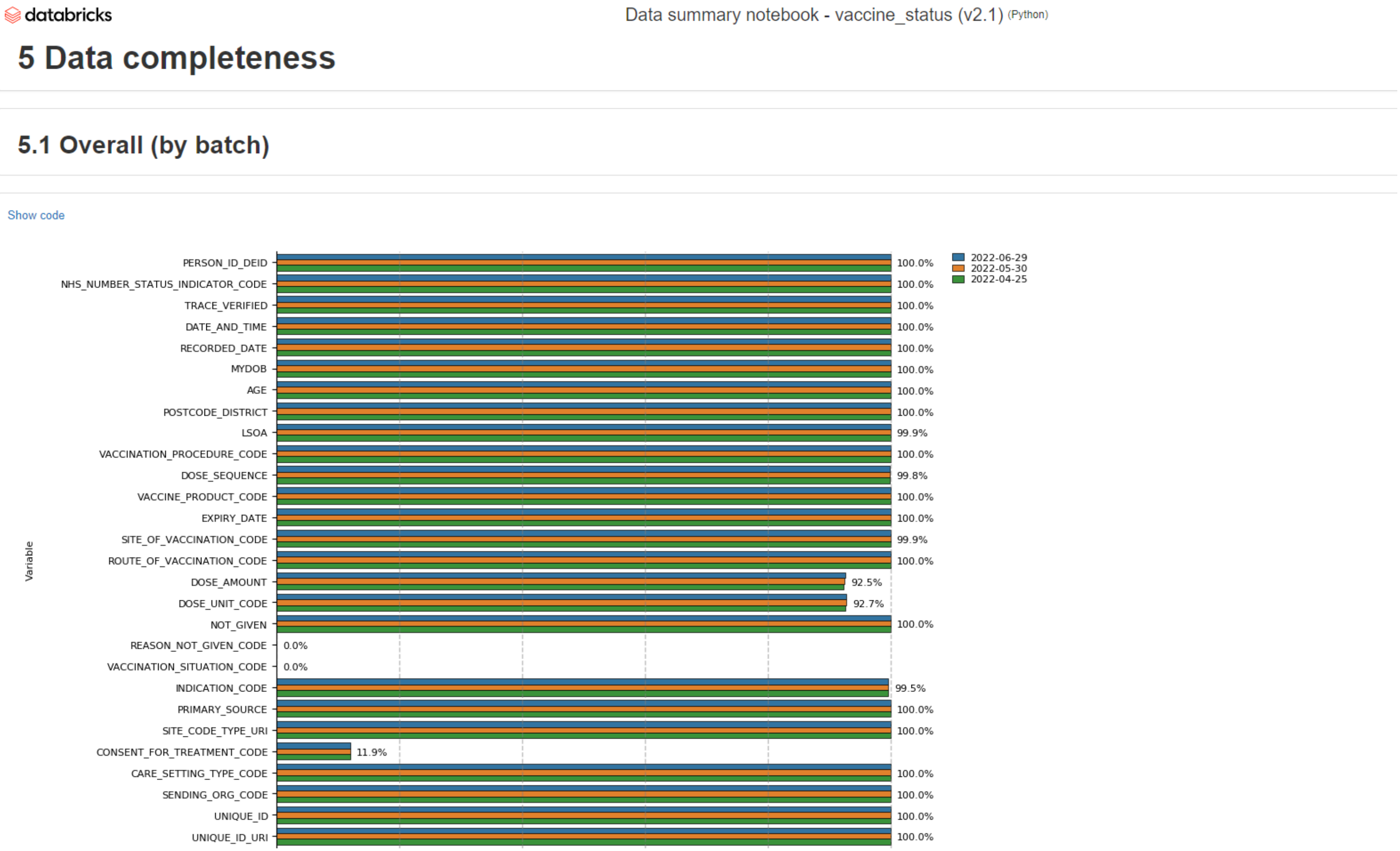
> Data insight notebooks

Code

> Demos

> Curated data

> Data curation pipeline functions



databricks

Data summary notebook - vaccine_status (v2.1) (Python)

5 Data completeness

5.1 Overall (by batch)

Show code

Resources

Data

> Data documentation

> Data dictionary

> Data summary notebooks

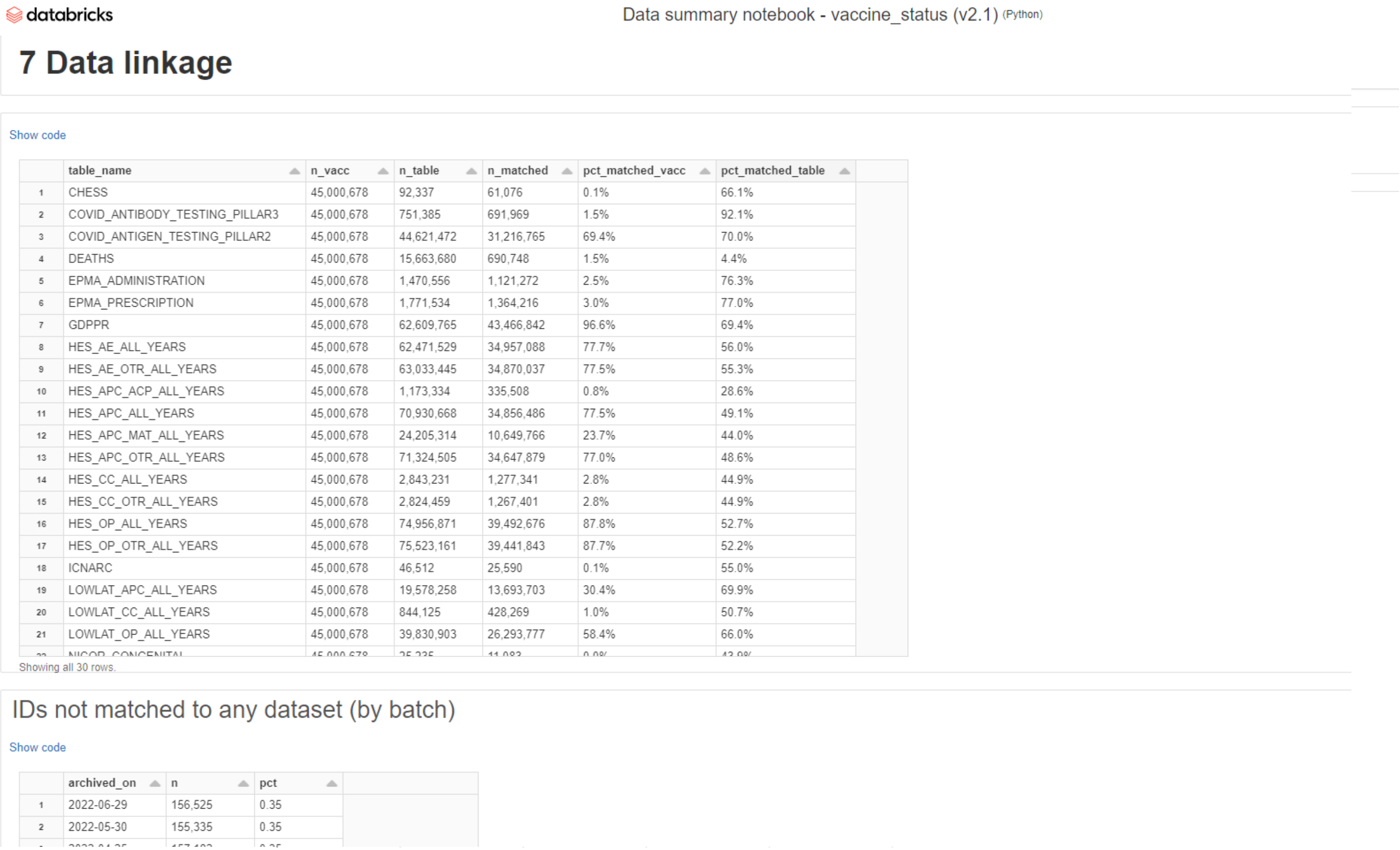
> Data insight notebooks

Code

> Demos

> Curated data

> Data curation pipeline functions



databricks

Data summary notebook - vaccine_status (v2.1) (Python)

7 Data linkage

Show code

	table_name	n_vacc	n_table	n_matched	pct_matched_vacc	pct_matched_table
1	CHESS	45,000,678	92,337	61,076	0.1%	66.1%
2	COVID_ANTIBODY_TESTING_PILLAR3	45,000,678	751,385	691,969	1.5%	92.1%
3	COVID_ANTIGEN_TESTING_PILLAR2	45,000,678	44,621,472	31,216,765	69.4%	70.0%
4	DEATHS	45,000,678	15,663,680	690,748	1.5%	4.4%
5	EPMA_ADMINISTRATION	45,000,678	1,470,556	1,121,272	2.5%	76.3%
6	EPMA_PRESCRIPTION	45,000,678	1,771,534	1,364,216	3.0%	77.0%
7	GDPPR	45,000,678	62,609,765	43,466,842	96.6%	69.4%
8	HES_AE_ALL_YEARS	45,000,678	62,471,529	34,957,088	77.7%	56.0%
9	HES_AE_OTR_ALL_YEARS	45,000,678	63,033,445	34,870,037	77.5%	55.3%
10	HES_APC_ACP_ALL_YEARS	45,000,678	1,173,334	335,508	0.8%	28.6%
11	HES_APC_ALL_YEARS	45,000,678	70,930,668	34,856,486	77.5%	49.1%
12	HES_APC_MAT_ALL_YEARS	45,000,678	24,205,314	10,649,766	23.7%	44.0%
13	HES_APC_OTR_ALL_YEARS	45,000,678	71,324,505	34,647,879	77.0%	48.6%
14	HES_CC_ALL_YEARS	45,000,678	2,843,231	1,277,341	2.8%	44.9%
15	HES_CC_OTR_ALL_YEARS	45,000,678	2,824,459	1,267,401	2.8%	44.9%
16	HES_OP_ALL_YEARS	45,000,678	74,956,871	39,492,676	87.8%	52.7%
17	HES_OP_OTR_ALL_YEARS	45,000,678	75,523,161	39,441,843	87.7%	52.2%
18	ICNARC	45,000,678	46,512	25,590	0.1%	55.0%
19	LOWLAT_APC_ALL_YEARS	45,000,678	19,578,258	13,693,703	30.4%	69.9%
20	LOWLAT_CC_ALL_YEARS	45,000,678	844,125	428,269	1.0%	50.7%
21	LOWLAT_OP_ALL_YEARS	45,000,678	39,830,903	26,293,777	58.4%	66.0%
22	NICOR_CONGENITAL	45,000,678	25,235	11,083	0.0%	43.9%

Showing all 30 rows.

IDs not matched to any dataset (by batch)

Show code

	archived_on	n	pct
1	2022-06-29	156,525	0.35
2	2022-05-30	155,335	0.35
3	2022-04-25	[illegible]	0.35

Resources

Data

> Data documentation

> Data dictionary

> Data summary notebooks

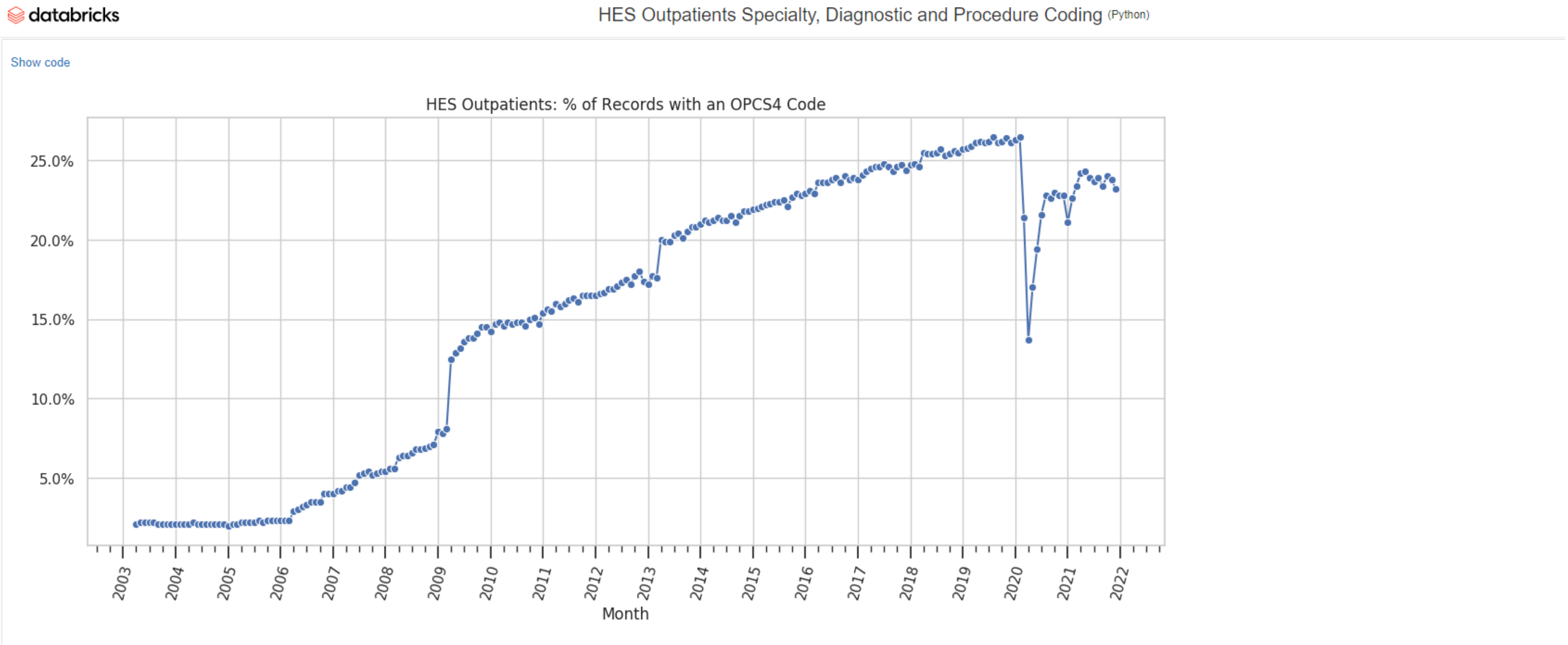
> Data insight notebooks

Code

> Demos

> Curated data

> Data curation pipeline functions



databricks

HES Outpatients Specialty, Diagnostic and Procedure Coding (Python)

Resources

Data

> Data documentation

> Data dictionary

> Data summary notebooks

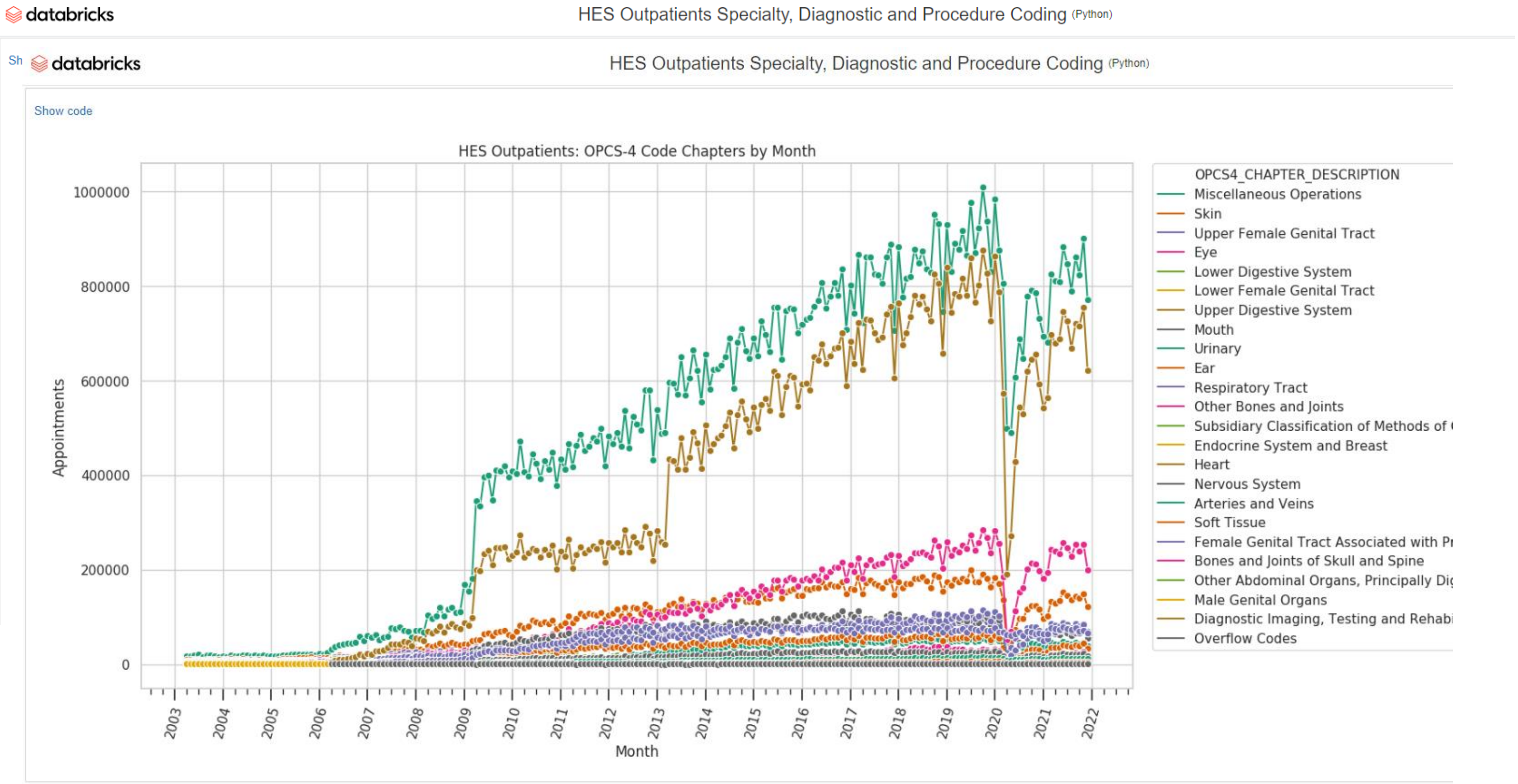
> Data insight notebooks

Code

> Demos

> Curated data

> Data curation pipeline functions


databricks HES Outpatients Specialty, Diagnostic and Procedure Coding (Python)

databricks HES Outpatients Specialty, Diagnostic and Procedure Coding (Python)

Show code

HES Outpatients: OPCS-4 Code Chapters by Month
Appointments
1000000
800000
600000
400000
200000
0
Month
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
OPCS4_CHAPTER_DESCRIPTION
Miscellaneous Operations
Skin
Upper Female Genital Tract
Eye
Lower Digestive System
Lower Female Genital Tract
Upper Digestive System
Mouth
Urinary
Ear
Respiratory Tract
Other Bones and Joints
Subsidiary Classification of Methods of
Endocrine System and Breast
Heart
Nervous System
Arteries and Veins
Soft Tissue
Female Genital Tract Associated with Pr
Bones and Joints of Skull and Spine
Other Abdominal Organs, Principally Di
Male Genital Organs
Diagnostic Imaging, Testing and Rehabi
Overflow Codes

Resources

Data

> Data documentation

> Data dictionary

> Data summary notebooks

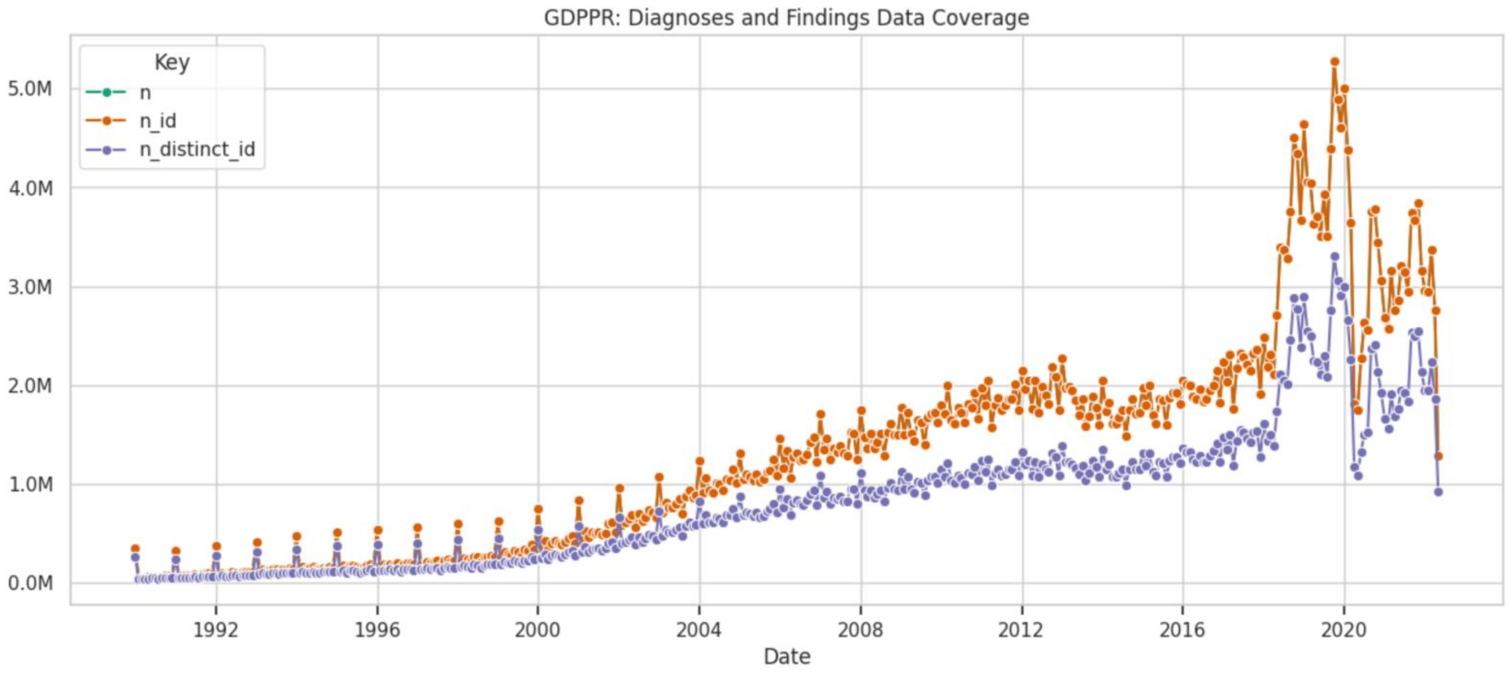
> Data insight notebooks

Code

> Demos

> Curated data

> Data curation pipeline functions



Resources

Data

> Data documentation

> Data dictionary

> Data summary notebooks

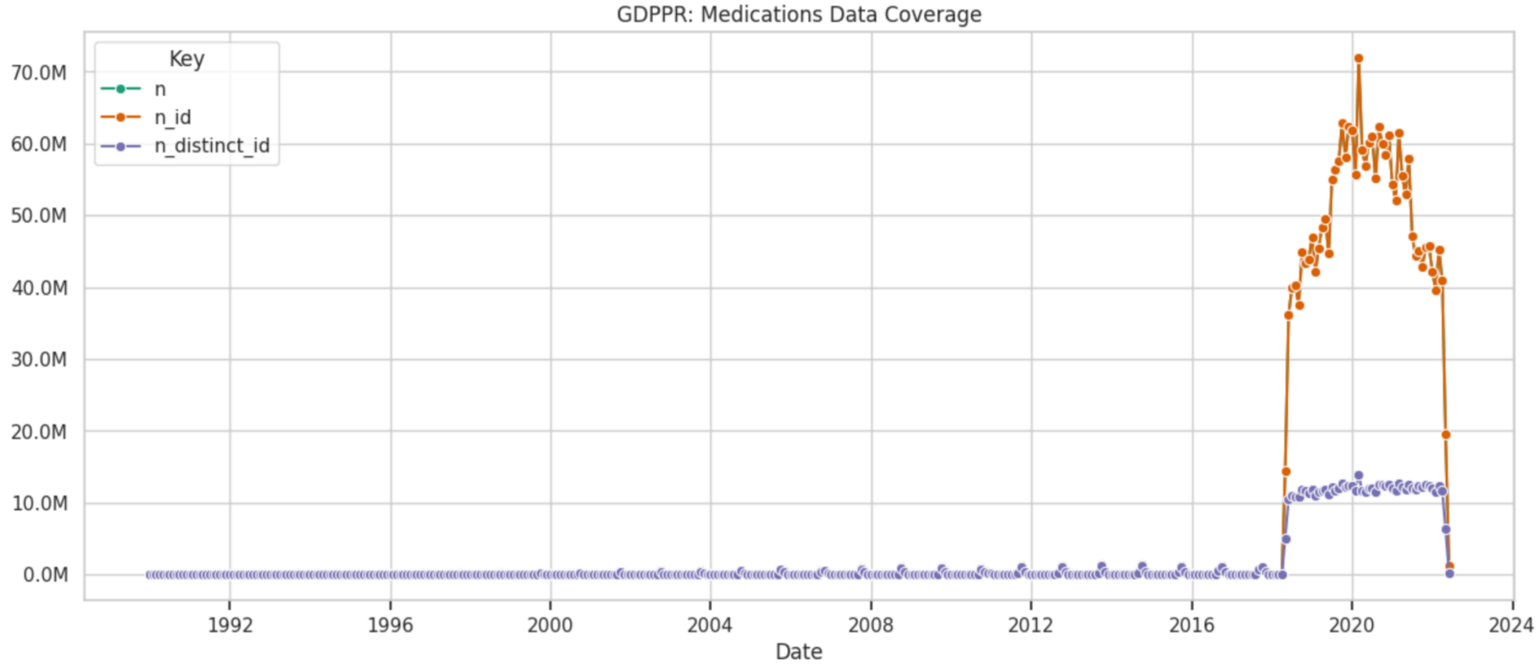
> Data insight notebooks

Code

> Demos

> Curated data

> Data curation pipeline functions



British Heart Foundation Data Science Centre
Led by Health Data Research UK

Resources

Data

> Data documentation

> Data dictionary

> Data summary notebooks

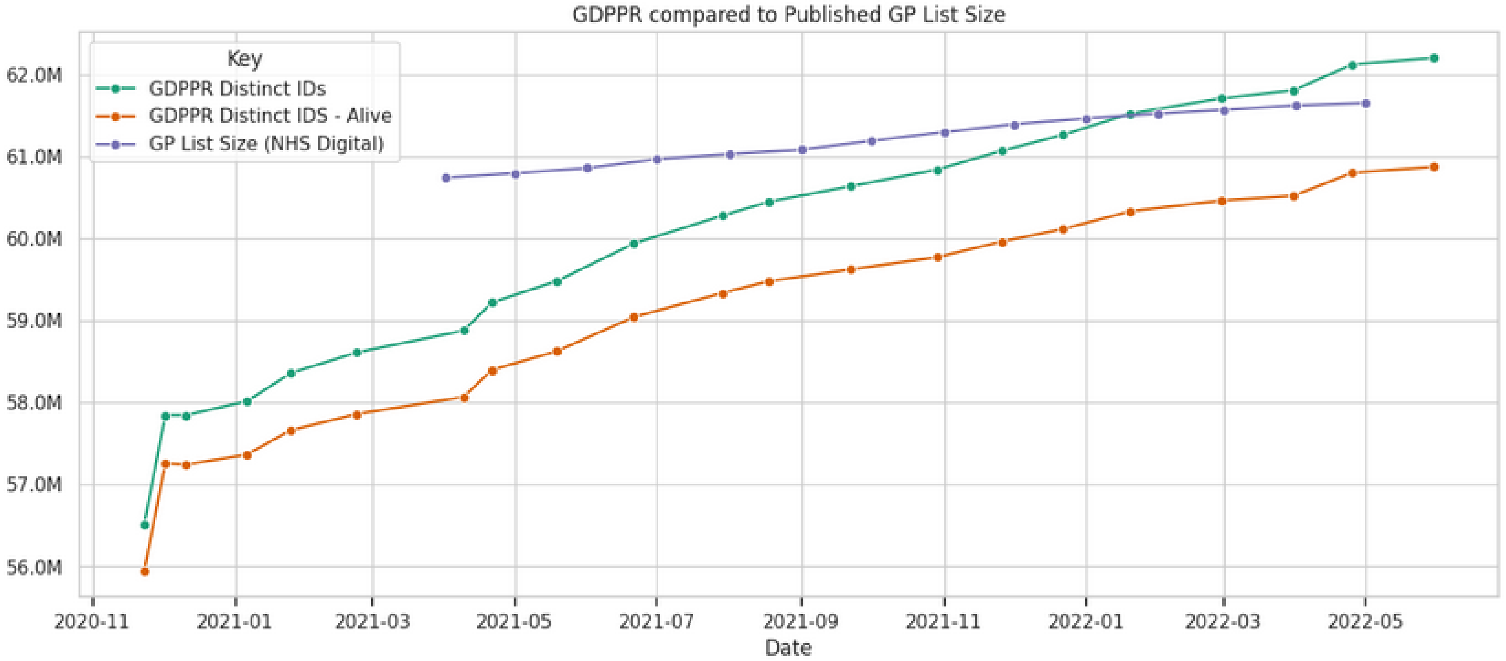
> Data insight notebooks

Code

> Demos

> Curated data

> Data curation pipeline functions



Resources

Data

> Data documentation

> Data dictionary

> Data summary notebooks

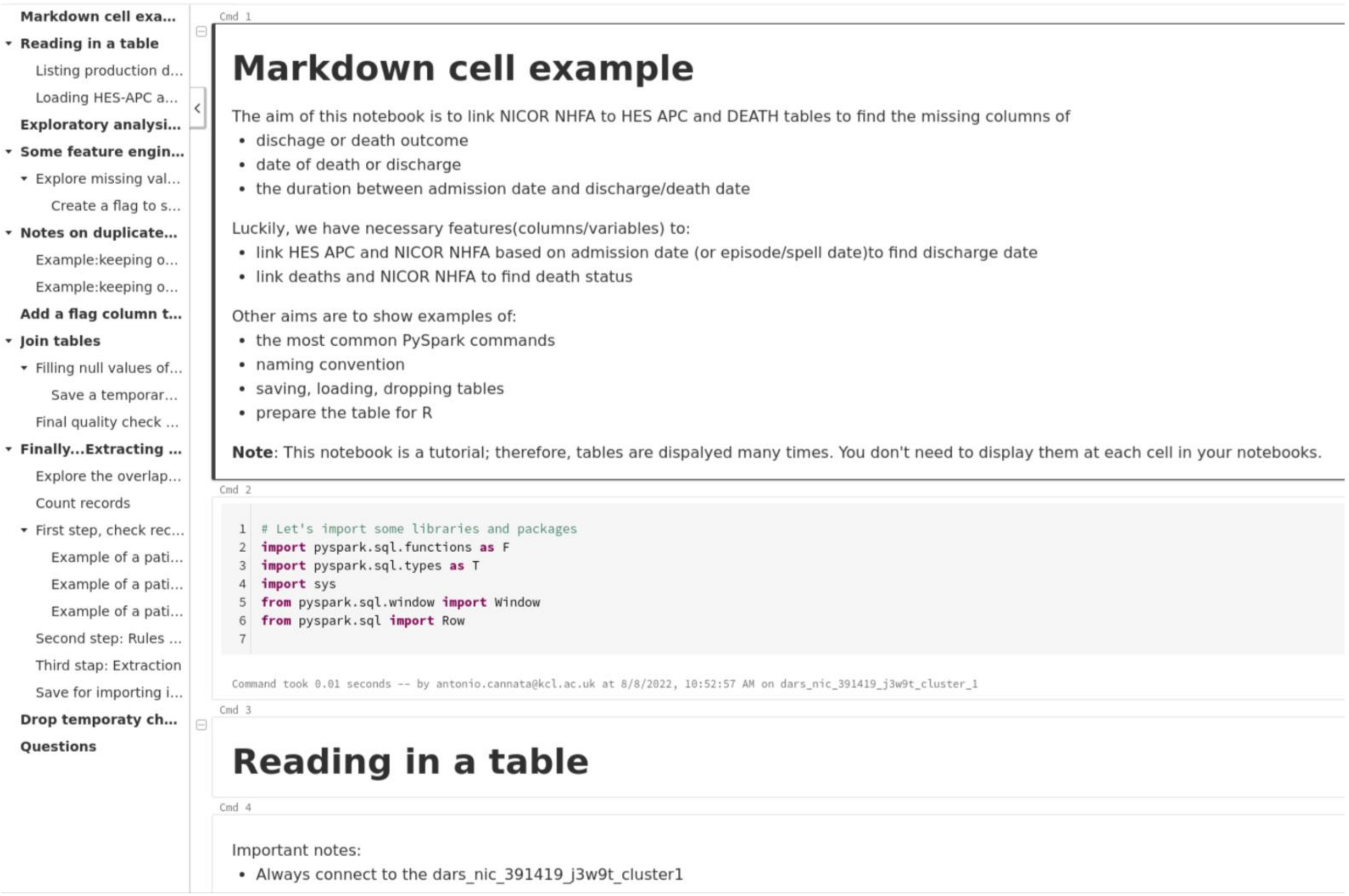
> Data insight notebooks

Code

> Demos

> Curated data

> Data curation pipeline functions



Markdown cell example

The aim of this notebook is to link NICOR NHFA to HES APC and DEATH tables to find the missing columns of

- dischage or death outcome
- date of death or discharge
- the duration between admission date and discharge/death date

Luckily, we have necessary features(columns/variables) to:

- link HES APC and NICOR NHFA based on admission date (or episode/spell date)to find discharge date
- link deaths and NICOR NHFA to find death status

Other aims are to show examples of:

- the most common PySpark commands
- naming convention
- saving, loading, dropping tables
- prepare the table for R

Note: This notebook is a tutorial; therefore, tables are dispalyed many times. You don't need to display them at each cell in your notebooks.

```
# Let's import some libraries and packages
import pyspark.sql.functions as F
import pyspark.sql.types as T
import sys
from pyspark.sql.window import Window
from pyspark.sql import Row

```

```
Command took 0.01 seconds -- by antonio.cannata@kcl.ac.uk at 8/8/2022, 10:52:57 AM on dars_nic_391419_j3w9t_cluster_1
```

Reading in a table

Important notes:

- Always connect to the dars_nic_391419_j3w9t_cluster1

Resources

Data

> Data documentation

> Data dictionary

> Data summary notebooks

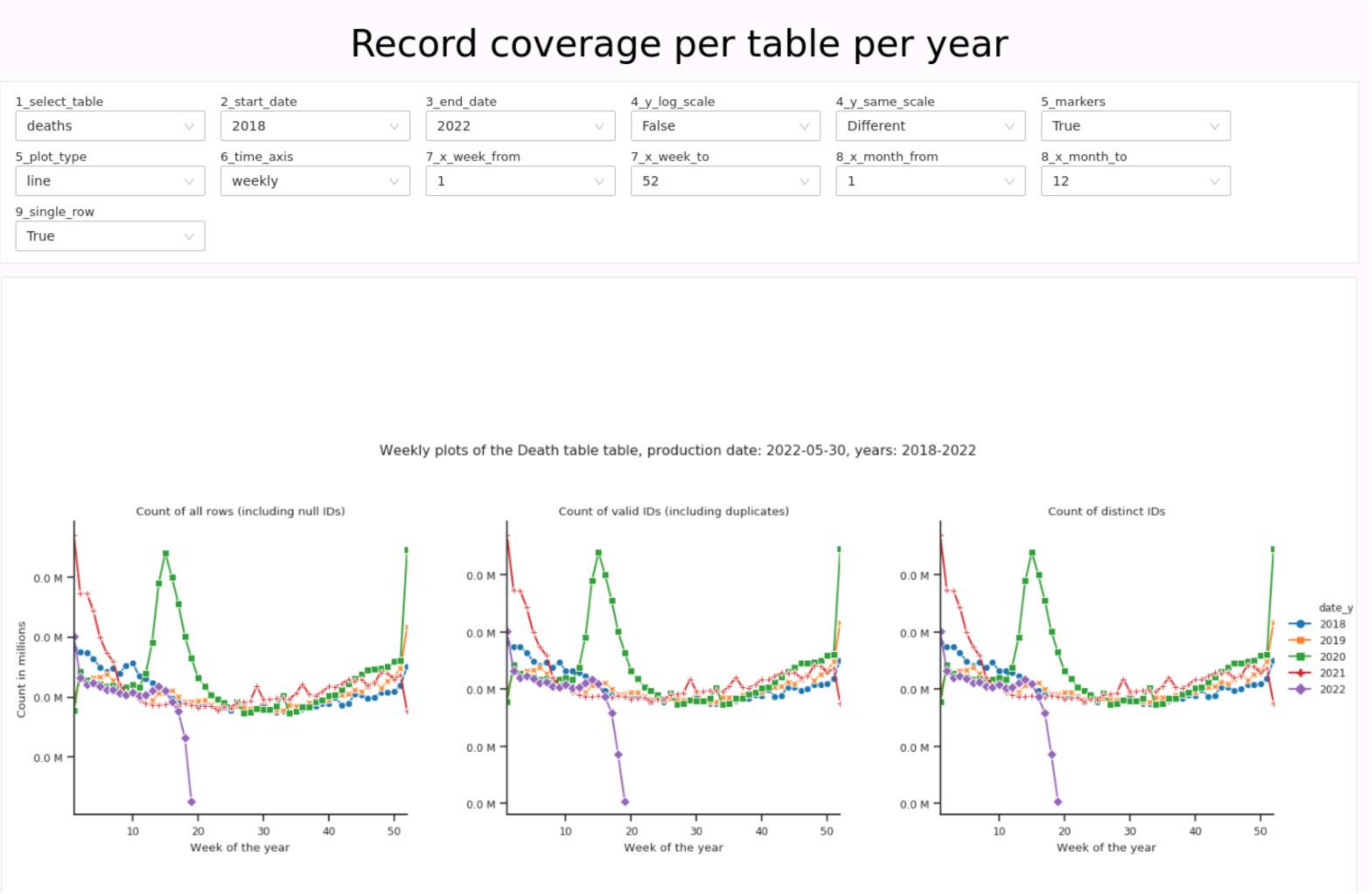
> Data insight notebooks

Code

> Demos

> Curated data

> Data curation pipeline functions



Record coverage per table per year

1_select_table	2_start_date	3_end_date	4_y_log_scale	4_y_same_scale	5_markers
deaths	2018	2022	False	Different	True
5_plot_type	6_time_axis	7_x_week_from	7_x_week_to	8_x_month_from	8_x_month_to
line	weekly	1	52	1	12
9_single_row					
True					

Weekly plots of the Death table table, production date: 2022-05-30, years: 2018-2022

Resources

Data

> Data documentation

> Data dictionary

> Data summary notebooks

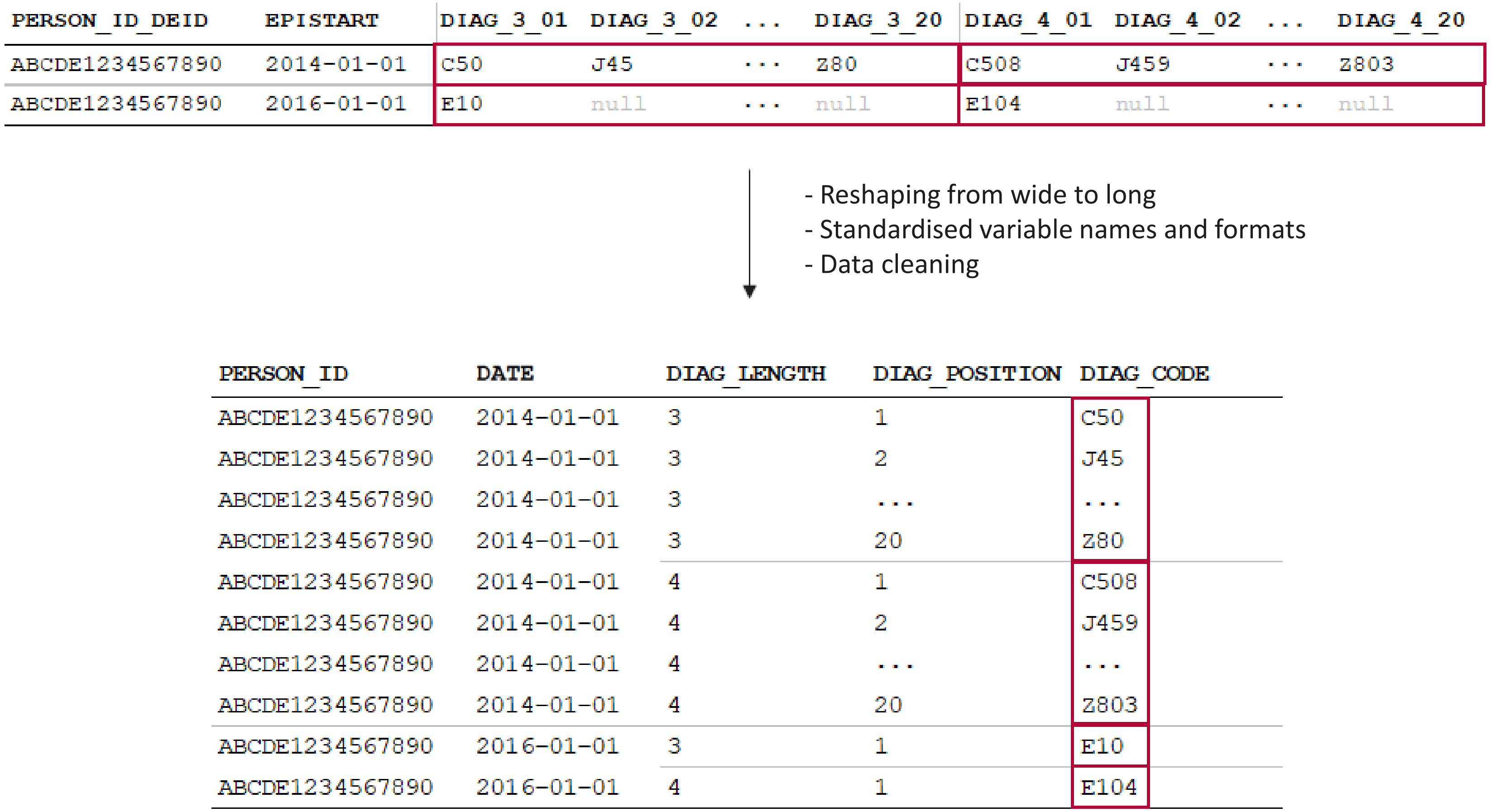
> Data insight notebooks

Code

> Demos

> Curated data

> Data curation pipeline functions



PERSON_ID_DEID	EPISTART	DIAG_3_01	DIAG_3_02	...	DIAG_3_20	DIAG_4_01	DIAG_4_02	...	DIAG_4_20
ABCDE1234567890	2014-01-01	C50	J45	...	Z80	C508	J459	...	Z803
ABCDE1234567890	2016-01-01	E10	null	...	null	E104	null	...	null

- Reshaping from wide to long
- Standardised variable names and formats
- Data cleaning

PERSON_ID	DATE	DIAG_LENGTH	DIAG_POSITION	DIAG_CODE
ABCDE1234567890	2014-01-01	3	1	C50
ABCDE1234567890	2014-01-01	3	2	J45
ABCDE1234567890	2014-01-01	3	...	...
ABCDE1234567890	2014-01-01	3	20	Z80
ABCDE1234567890	2014-01-01	4	1	C508
ABCDE1234567890	2014-01-01	4	2	J459
ABCDE1234567890	2014-01-01	4	...	...
ABCDE1234567890	2014-01-01	4	20	Z803
ABCDE1234567890	2016-01-01	3	1	E10
ABCDE1234567890	2016-01-01	4	1	E104

Resources

Data

> Data documentation

> Data dictionary

> Data summary notebooks

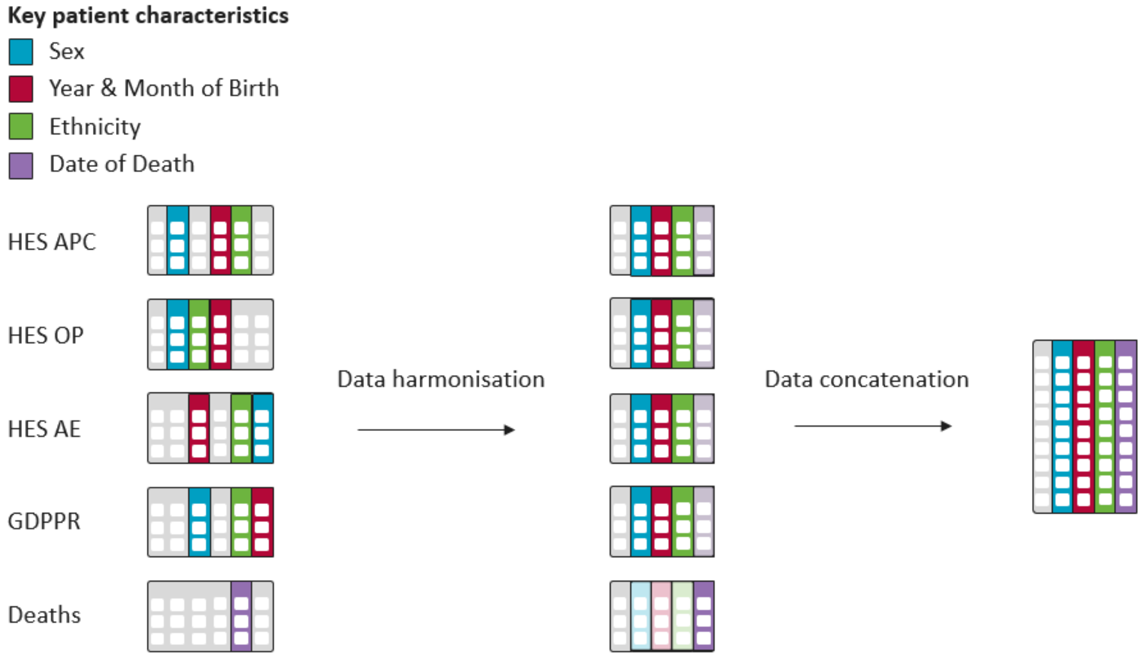
> Data insight notebooks

Code

> Demos

> Curated data

> Data curation pipeline functions



Resources

Data

> Data documentation

> Data dictionary

> Data summary notebooks

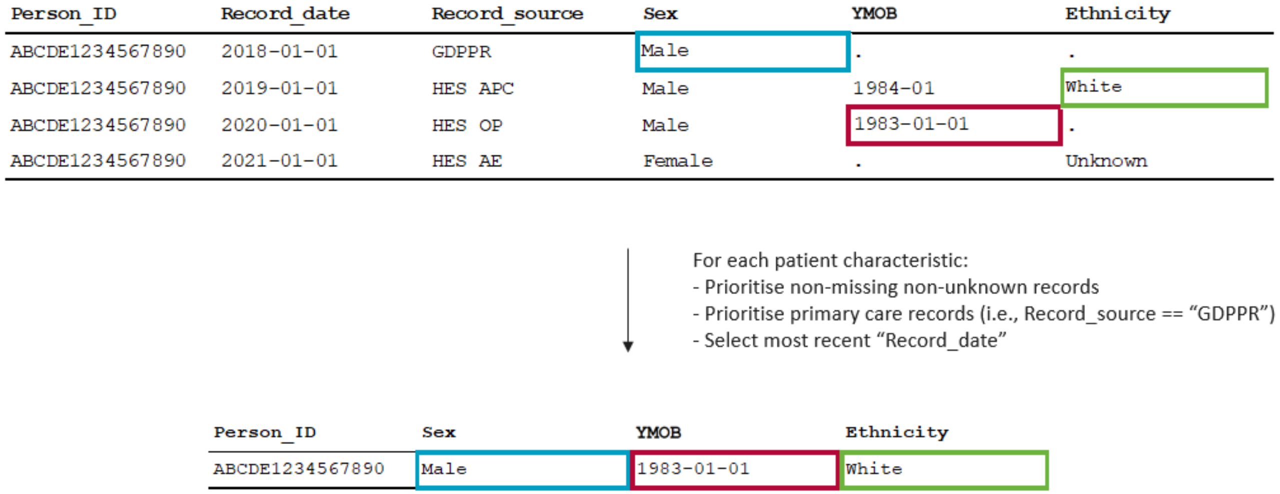
> Data insight notebooks

Code

> Demos

> Curated data

> Data curation pipeline functions



Person_ID	Record_date	Record_source	Sex	YMOB	Ethnicity
ABCDE1234567890	2018-01-01	GDPPR	Male	.	.
ABCDE1234567890	2019-01-01	HES APC	Male	1984-01	White
ABCDE1234567890	2020-01-01	HES OP	Male	1983-01-01	.
ABCDE1234567890	2021-01-01	HES AE	Female	.	Unknown

For each patient characteristic:
- Prioritise non-missing non-unknown records
- Prioritise primary care records (i.e., Record_source == "GDPPR")
- Select most recent "Record_date"

Person_ID	Sex	YMOB	Ethnicity
ABCDE1234567890	Male	1983-01-01	White

Resources

Data

> Data documentation

> Data dictionary

> Data summary notebooks

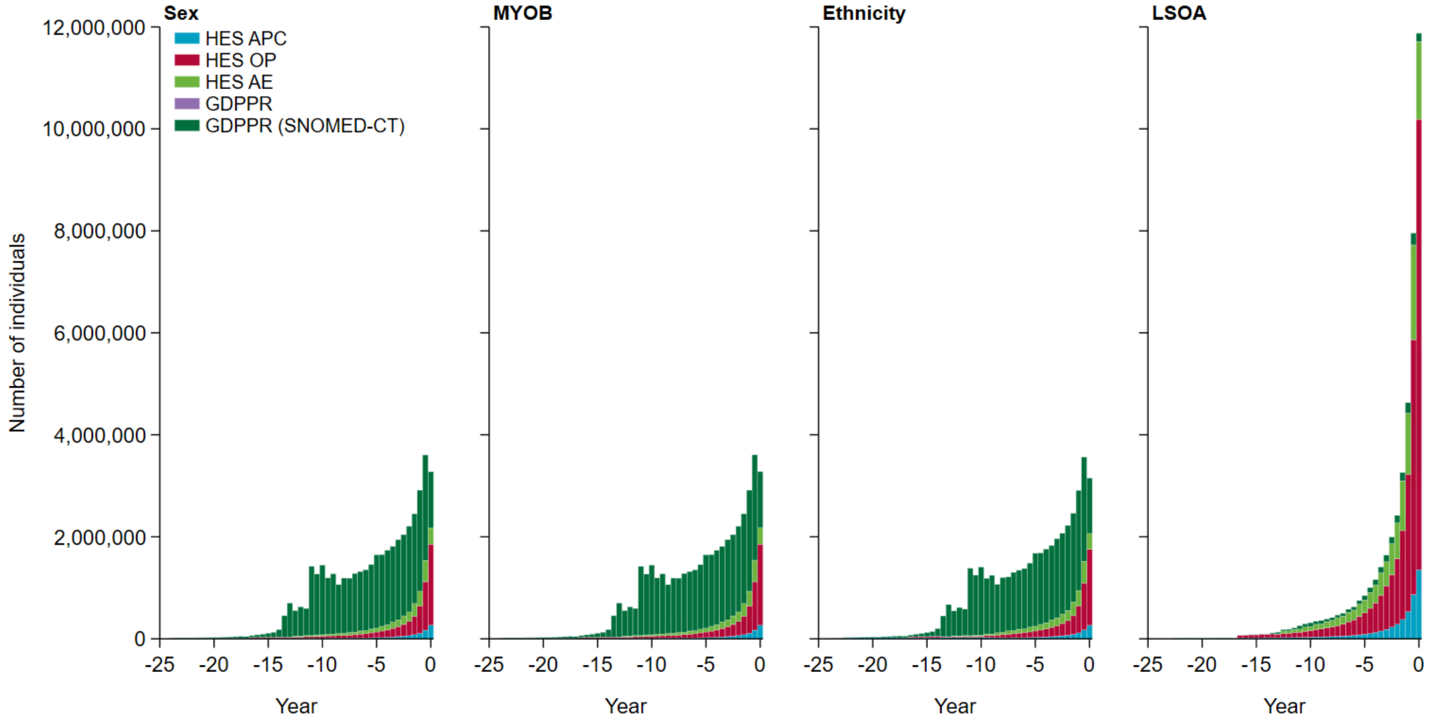
> Data insight notebooks

Code

> Demos

> Curated data

> Data curation pipeline functions



Resources

Data

> Data documentation

> Data dictionary

> Data summary notebooks

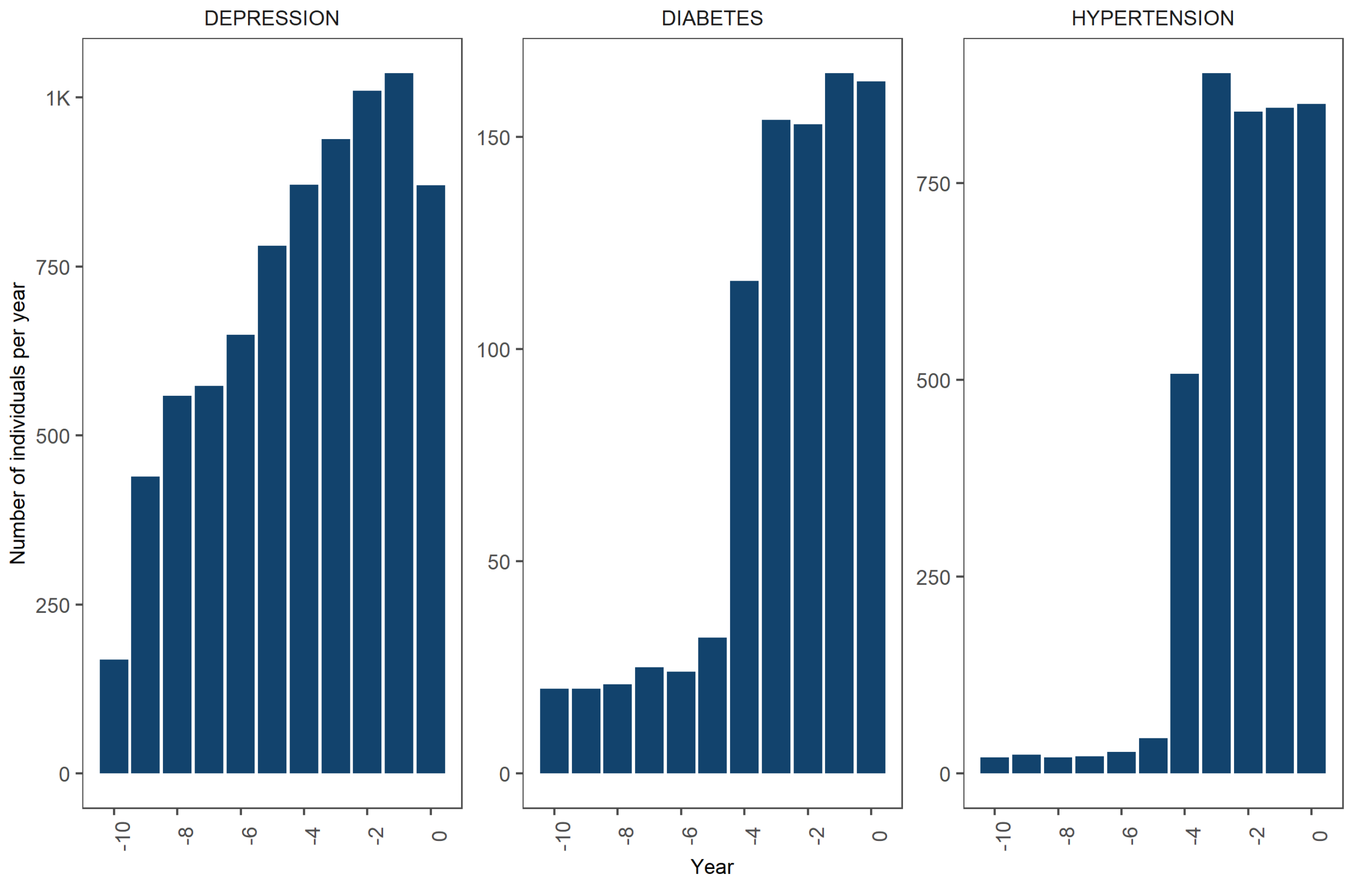
> Data insight notebooks

Code

> Demos

> Curated data

> Data curation pipeline functions



Resources

Data

> Data documentation

> Data dictionary

> Data summary notebooks

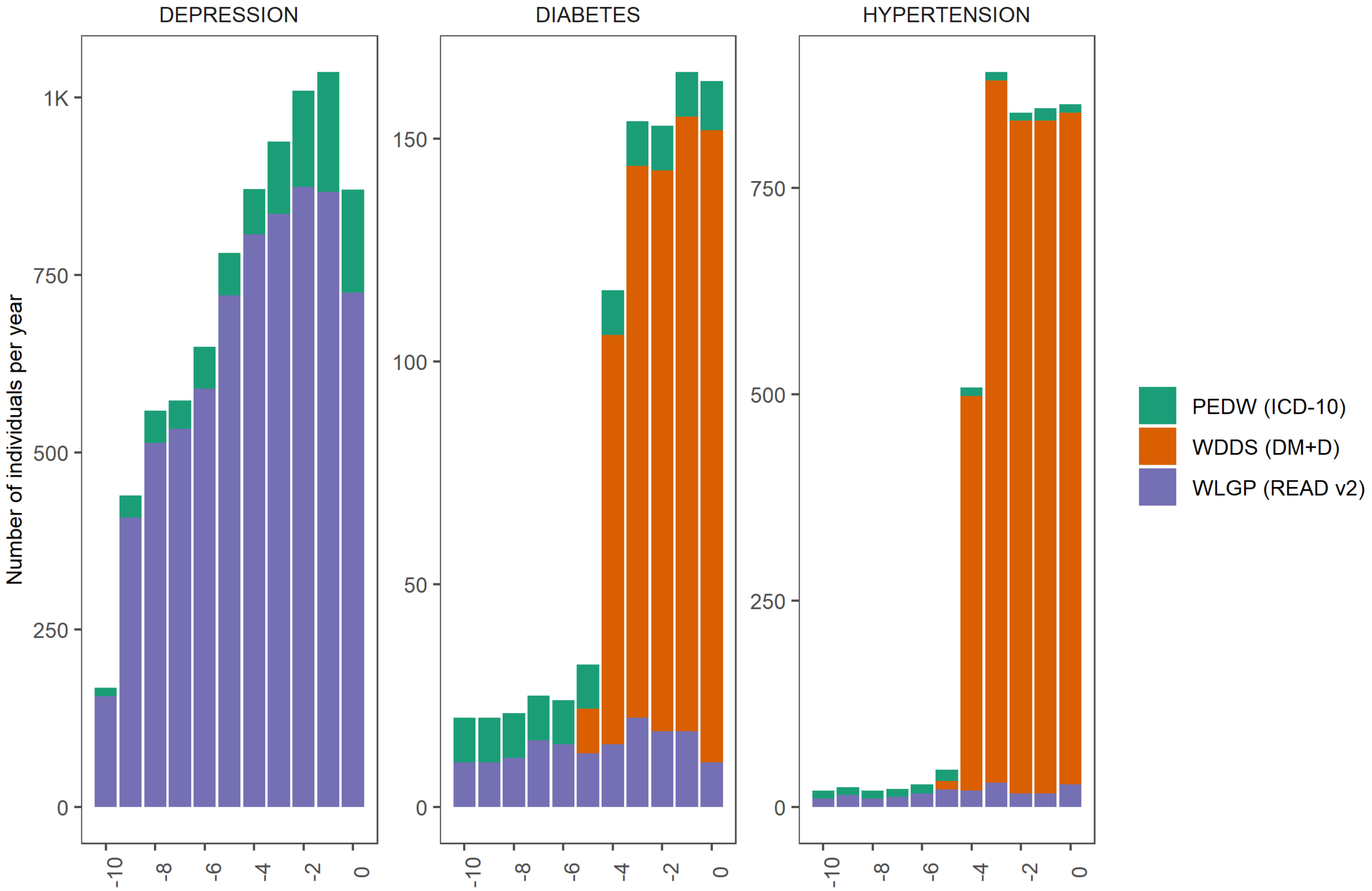
> Data insight notebooks

Code

> Demos

> Curated data

> Data curation pipeline functions



Resources

Data

> Data documentation

> Data dictionary

> Data summary notebooks

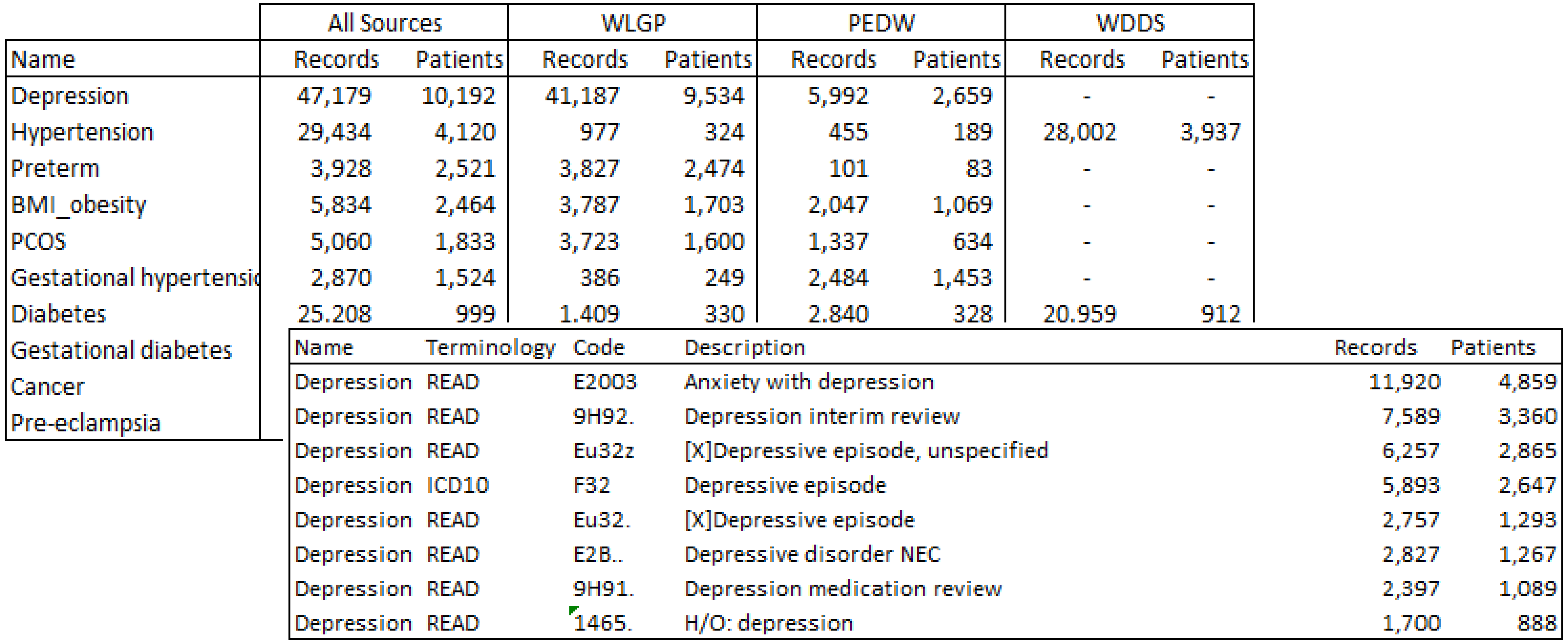
> Data insight notebooks

Code

> Demos

> Curated data

> Data curation pipeline functions



	All Sources		WLGP		PEDW		WDDS	
Name	Records	Patients	Records	Patients	Records	Patients	Records	Patients
Depression	47,179	10,192	41,187	9,534	5,992	2,659	-	-
Hypertension	29,434	4,120	977	324	455	189	28,002	3,937
Preterm	3,928	2,521	3,827	2,474	101	83	-	-
BMI_obesity	5,834	2,464	3,787	1,703	2,047	1,069	-	-
PCOS	5,060	1,833	3,723	1,600	1,337	634	-	-
Gestational hypertensio	2,870	1,524	386	249	2,484	1,453	-	-
Diabetes	25,208	999	1,409	330	2,840	328	20,959	912
Gestational diabetes								
Cancer								
Pre-eclampsia								

Name	Terminology	Code	Description	Records	Patients
Depression	READ	E2003	Anxiety with depression	11,920	4,859
Depression	READ	9H92.	Depression interim review	7,589	3,360
Depression	READ	Eu32z	[X]Depressive episode, unspecified	6,257	2,865
Depression	ICD10	F32	Depressive episode	5,893	2,647
Depression	READ	Eu32.	[X]Depressive episode	2,757	1,293
Depression	READ	E2B..	Depressive disorder NEC	2,827	1,267
Depression	READ	9H91.	Depression medication review	2,397	1,089
Depression	READ	1465.	H/O: depression	1,700	888

Project Support

Health Data Science Team

- Review project proposals
- Understand project requirements
- Signpost to:
  - Data resources
  - Demos
  - Reusable/adaptable code
- Development:
  - Data curation pipelines

Stages of a data curation pipeline

> Parameters

> Code-list

> Cohort selection

> Data freezing / snapshots

> Data cleaning / reformatting

> Key patient characteristics

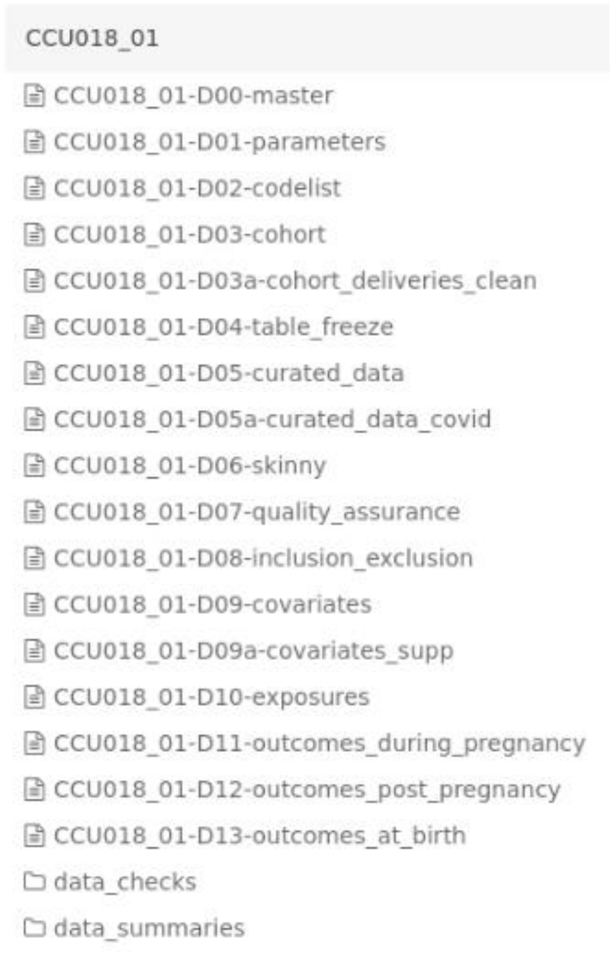
> Quality assurance

> Inclusion / exclusion

> Covariates

> Exposures

> Outcomes



CCU018_01

- CCU018_01-D00-master
- CCU018_01-D01-parameters
- CCU018_01-D02-codelist
- CCU018_01-D03-cohort
- CCU018_01-D03a-cohort_deliveries_clean
- CCU018_01-D04-table_freeze
- CCU018_01-D05-curated_data
- CCU018_01-D05a-curated_data_covid
- CCU018_01-D06-skinny
- CCU018_01-D07-quality_assurance
- CCU018_01-D08-inclusion_exclusion
- CCU018_01-D09-covariates
- CCU018_01-D09a-covariates_supp
- CCU018_01-D10-exposures
- CCU018_01-D11-outcomes_during_pregnancy
- CCU018_01-D12-outcomes_post_pregnancy
- CCU018_01-D13-outcomes_at_birth
- data_checks
- data_summaries

Thank you for listening

john.nolan@hdruk.ac.uk